**REID (Research and Evaluation in Education), 10(1), 2024, 114-127**
Available online at: http://journal.uny.ac.id/index.php/reid

# Stability of estimation item parameter in IRT dichotomy considering the number of participants

**Zulfa Safina Ibrahim[1]*; Heri Retnawati[1]; Alfred Irambona[2]; Beatriz Eugenia Orantes Pérez[3]**
[1]Universitas Negeri Yogyakarta, Indonesia
[2]Burundi University, Burundi
[3]El Colegio de la frontera sur (ECOSUR), Mexico
*Corresponding Author. E-mail: zulfasafina.2022@student.uny.ac.id

**ARTICLE INFO**

**Article History**
**Submitted:**
10 May 2024
**Revised:**
29 June 2024
**Accepted:**
30 June 2024

**Keywords**
stability; item parameter estimation; item response theory; EAP; bootstrapping

**Scan Me:**

**ABSTRACT**

This research is related to item response theory (IRT) which is needed to measure the goodness of a test set, while item parameter estimation is needed to determine the technical properties of a test item. Stability of item parameter estimation is conducted to determine the minimum sample that can be used to obtain good item parameter estimation results. The purpose of this study is to describe the effect of the number of test takers on the stability of item parameter estimation with the Bayes method (expected a posteriori, EAP) on dichotomous data. This research is an exploratory descriptive research with a bootstrap approach using the EAP method. The EAP method is performed by modifying the likelihood and function to include prior information about the participant's 9 score. Bootstrapping on the original data is done to take bootstrap samples. with ten different sample sizes of 100, 150, 250, 300, 500, 700, 1,000, 1,500, 2,000, 2,500 were then replicated ten times and grain parameter estimation was performed. Each sample data with ten replications was calculated Root Mean Squared Difference (RMSD) value. The results showed that the 2PL model was chosen as the best model. The RMSD value obtained proves that many test participants affect the stability of item parameter estimation on dichotomous data with the 2PL model. The minimum sample to ensure the stability of item parameter estimates with the 2PL model is 1,000 test participants.



**To cite this article (in APA style):**
Ibrahim, Z., Retnawati, H., Irambona, A., & Pérez, B. (2024). Stability of estimation item parameter in IRT dichotomy considering the number of participants. *REID (Research and Evaluation in Education), 10*(1), 114-127. doi:https://doi.org/10.21831/reid.v10i1.73055

## INTRODUCTION

Many language proficiency tests are conducted to measure a person's ability related to how deep a person's ability to a particular language (Gong et al., 2024; Neiriz, 2023). The language proficiency test device is made as a reference to measure the ability of participants, but it is not certain how well the test device measures the ability of test takers (Kim et al., 2023; Khodi et al., 2024; Sarac & Loken, 2023; Crivelli et al., 2021). Item parameter estimation is needed to determine the technical properties of a test item. Because the actual value of the item parameters on the test cannot be known, it is necessary to estimate them first. Then, after parameter estimation, information related to the technical properties of the test items will be obtained. The main purpose of giving tests to test takers in item response theory (IRT) is to determine the scale of the participant's ability. If a measure of ability has been obtained for each test taker, then test takers can be evaluated in terms of how much basic ability they have and comparisons between test takers can be made (Baker, 2001; Schleicher et al., 2017; Sabitova, 2023). Meanwhile, to estimate the item parameters on a device there are several variables that need to be considered, one of

which is the number of test participants. This variable is thought to affect the stability of the item parameter estimation, so it is necessary to conduct research on the stability of parameter estimation. According to Hambleton (1989), it is difficult to accurately determine the test length and sample size required for item parameter estimation in IRT. This study was conducted with IRT on dichotomous data using Bayes method (expected a posteriori, EAP) and Bootstrapping method. EAP method is used to calculate the estimate of the ability parameter (6). Estimating the value of 8 can be Maximum Likelihood Estimation (MLE) and Maximum a Posteriori (MAP) are two other methods. The basic concept of the MLE method is based on the maximum likelihood function. The weakness of the MLE method is that it is difficult to use. It is used to estimate 9 if the participant answers all right or all wrong. Thus, Hambleton et al. (1991) suggested using the Bayes estimation method. The EAP method is done by modifying the likelihood function to include prior information about participants' 9 values (Desjardins & Bulut, 2018). Bootstrapping is a type of database simulation that involves resampling data many times to produce empirical estimates of the entire statistical sampling distribution. Davidson and MacKinnon (1993) define that the idea of Bootstrapping is to use a set of available data to design a Monte Carlo-like experiment, where the data is used to estimate the error obtained from the empirical distribution function of the resulting samples. This study sampled data from the original data with returns and replicated 10 times, as based on the opinion of Harwell et al. (1996) that simulation-related research on IRT only requires a small number of replications of at least 10 times.

Research related to the stability of parameter estimation with Rasch, 1PL, and 3PL models has been conducted. One of them is conducted by Şahin and Anıl (2017) which concludes that there is a relationship between the length of the test and the number of test participants selected on the estimation of the item parameters as the number of items on the test. Another study by Susongko (2021) shows that the Rasch model has better stability in estimating participant ability compared to the 1PL model, as indicated by lower bias and standard error in scientific literacy tests. Meanwhile, Guo et al. (2021) reveal that although the 1PL model is simpler, it struggles with stability, especially with small sample sizes, where Bayesian modal estimation methods have been proposed to improve accuracy. Additionally, Falani et al. (2018) found that the 1PL, 2PL, and 3PL models each provide optimal parameter estimates when aligned with their respective characteristics, highlighting the importance of model selection for more accurate estimation.

Each model in 1PL, 2PL, and 3PL has estimation stability with different test lengths and number of tests. Research conducted by Stone and Yumoto (2004) related to the stability of item parameter estimates using the Rasch model or the 1PL model found that sample size is the main factor in obtaining stable item parameter estimates. Research using the Rasch model only focuses on item difficulty parameters. In the parameter estimation that has been done, a sample size of 500 is the minimum sample size for item parameter estimation to be accepted and said to be good. Therefore, researchers examined the stability of item parameter estimates with other models, namely the 2PL model or the 3PL model in IRT. If there is evidence that the number of test participants affects the item parameters, it can be concluded that the variable number of test participants affects the stability of the item parameter estimates. From the research results obtained, it is hoped that they can be used to help in the consideration of making good questions. This research is related to item response theory which is needed to see how well a test device is tested. Therefore, research to determine the stability of item parameter estimates on dichotomous data item response theory by considering the number of test participants needs to be done.

## METHOD

### Data Description

This study uses a bootstrap approach aiming to investigate the stability of item parameter estimates by considering the number of test participants using IRT dichotomous data. The data

used in this study are secondary data for the 2021 preTOEFL English/listening test device at one of the universities in Yogyakarta. The data is dichotomous scoring data from the preTOEFL English/listening test device. Population data on this study is 3,042 responses from test takers working on one of the identities (id) questions measuring preTOEFL English Listening skills.

Ethical clearance was obtained to conduct the research under the reference number T/1.1/UN34.9/KP.06.07/2022, issued by the Research Ethics Committee of Universitas Negeri Yogyakarta. The research participants were anonymous, and the data obtained in the study were guaranteed to be used only for research purposes.

## Data Analysis Steps

Data analysis was carried out in the following stages. (1) Preparing language proficiency test data. (2) Testing the fit of the best-fit model to the data using the Rasch model, 1PL model, 2PL model, or 3PL model using the Chi-Square test statistic. There are three models that can be used to conduct analysis using item response theory (Retnawati, 2014). The three models are distinguished from the characteristics of the item used. According to Hambleton et al. (1991), the value of O is in the range of [-00. o], while according to Retnawati (2014), the O value is in the range of [-4.4]. The value of a, is in the range of [0.2], b, is in the range of [-2.2], and c, is in the range of [0.1]. The models that can be used in IRT analysis are as follows, in which $i$ = 1,2,3, …, $m$; $j$ = 1,2,3, …, $n$; $m$ = number of items in the test; $n$ = number of participants in the test; $P(X_{ij} = 1)$ = probability of the *j-th* participant answering correctly on the item *i-th*; $b_i$= item difficulty of parameter i; $e$ = constants with value ranging from 2.718; D = constants with value ranging from 1.7; $\theta_j$ = *j-th* participant ability parameter; $a_i$ = a constant which is the level of item discriminating power (item discrimination) on the *i-th* item; and $c_i$ = item i's pseudo guessing parameter.

### *Rasch Model*

The Rasch model predicts the probability of correct answers using only one parameter, the item difficulty parameter (b). The item characteristic curve (ICC) equation for the Rasch model is presented in Equation (1).

$$P(X_{ij} = 1|\theta_j) = \frac{e^{(\theta_j - b_i)}}{1+e^{(\theta_j - b_i)}}, \quad \text{..... (1)}$$

### *1PL Model*

One of the most widely used item response theory models is the one-parameter Logistic model. The item characteristic curve equation for the 1PL model is presented in Equation (2).

$$P(X_{ij} = 1|\theta_j) = \frac{e^{(Da\{\theta_j - b_i\})}}{1+e^{(a\{\theta_j - b_i\})}}, \quad \text{..... (2)}$$

### *2PL Model*

The 2PL model predicts the probability of a correct answer using two parameters, namely the item difficulty parameter (b) and the item distinctiveness parameter (a). The ICC equation for the 2PL model is presented in Equation (3).

$$P(X_{ij} = 1|\theta_j) = \frac{e^{a_i(\theta_j - b_i)}}{1+e^{a_i(\theta_j - b_i)}}, \quad \text{..... (3)}$$

### *3PL Model*

The 3PL model is used to predict the probability of a participant's response answering correctly as in the 1PL and 2PL models but is constrained by a third parameter, the pseudo guessing

parameter (c), which limits the probability of supporting a participant's response answering correctly when the participant answers correctly. The participant's ability is close to $-\infty$. The ICC equation for the 3PL model is shown in Equation (4).

$$P(X_{ij} = 1|\theta_j) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1+e^{a_i(\theta_j - b_i)}}, \quad \text{..... (4)}$$

Then, the data analysis steps proceed to the following stages. (3) Testing IRT assumptions: unidimensionality, parameter invariance, and local independence. Each is elaborated as follows.

Unidimensionality means that each test item only measures one ability, and this assumption can only be demonstrated if the test contains one dominant component that can measure participants' achievements. One way to test the assumption of unidimensionality is to analyze the eigenvalue of the inter-item correlation matrix on the unidimensionality completion (Retnawati, 2014). The calculation of eigenvalues can be done using Principal Component Analysis (PCA) analysis and with the help of RStudio software using the PCA function and get_eigenvalue in the factoextra package (Kassambara & Mundt, 2020). If the unidimensionality assumption has been met, then the assumption of local independence has also been met (Hambleton et al., 1991). This assumption will be met if the answer between participants to an item does not affect the participant's answer to another item. This can be seen from the correlation value of the data using the cor() function in the RStudio sofiware. The assumption of parameter invariance means that item characteristics do not depend on the distribution of test-taker ability parameters. Conversely, the parameters that characterize test takers do not depend on item characteristics (Retnawati, 2014).

(4) Estimating item parameters with the Chi-Square test using Yen's method and on the original English/listening test device data based on the best model. There are two common ways to test model fit, namely, model fit test statistics and graphs. In the model fit statistics, Yen's QI method is a frequently used method. O, is one of the statistical tests of model fit used to test latent models. To calculate O, test takers were previously divided into 10 groups with the same number of members per group (Yen, 1981). The equation for Yen's Q1 method is shown in Equation (5), in which $i$ = number of items (i=1,2,3,...,n); $j$ = number of participant groups (j=1,2,3,...,10); $Q_{1i}$ = statistical value for item $i$; $N_j$ = number of participants in the j-th group; $O_{ij}$ = proportion of test-takers in the jth group who answered correctly on the item i-th; $E_{ij}$ = the predicted proportion of test-takers in the j-th group who answered correctly on the i-th item.

$$Q_{1i} = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1-E_{ij})}, \quad \text{..... (5)}$$

(5) Estimating ability parameters using the EAP method on the original English listening test device data based on the best model. The EAP method modifies the likelihood function to include prior information about the participants' 6 values. The EAP method is also a variation of the MAP method that uses the average 9 value from the posterior distribution (Desjardins & Bulut, 2018). According to Bock and Aitkin (1981), the calculation of the EAP method can be obtained by Equation (6), given that $x_j = \begin{cases} 1 \text{ correct answer} \\ 0 \text{ other} \end{cases}$, where $x_j$ is the score for item j and $P(k_j = 1|\theta) = \equiv_j (\theta)$ is the probability of answering correctly $x = 1$ a point $\theta$ on continuous ability. Furthermore, the likelihood of $\theta$ for a given response pattern $[x_1, x_2, \ldots, x_j]$ is as follows.

$$L_J(\theta) = \prod_{j=1}^{J} [\equiv_j (\theta)]^{x_j} [1 - \equiv_j (\theta)]^{1-x_j} \quad \text{..... (6)}$$

In the jth trial of the adaptive test, the provisional EAP estimate of the ability of ith participant, $\bar{\theta}_J$ can be approximated by Equation (7), and posterior standard deviation (PSD) can be approxi-

mated by Equation (8). In Equation (7) and Equation (8), $X_k$ is one of the g quadrature points and $W(X_k)$ is the weight associated with the point, which is normalized as in Equation (9).

$$\bar{\theta}_J = \frac{\sum_{k=1}^{q} X_k L_J(X_k) W(X_k)}{\sum_{k=1}^{q} L_J(X_k) W(X_k)}, \quad \text{(7)}$$

$$PSD(\theta) = \left[\frac{\sum_{k=1}^{q} (X_k - \bar{\theta}_J)^2 L_J(X_k) W(X_k)}{\sum_{k=1}^{q} L_J(X_k) W(X_k)}\right]^{\frac{1}{2}} \quad \text{(8)}$$

$$\sum_{k=1}^{q} W(X_k) = 1. \quad \text{(9)}$$

In the context of EAP estimation, weights are probabilities at corresponding points of the discrete prior distribution. In certain cases, e.g. when the normal prior distribution is summed, the points and weights can be chosen to improve the numerical accuracy of the integral estimate (Bock & Aitkin, 1981). This study uses the help of RStudio software to calculate $\theta$ using the EAP method. The functions provided by RStudio software to calculate $\theta$ are function mirt, coef, fscores, and itemfit in package mirt (Chalmers, 2012).

(6) Bootstrapping using an R program to generate data with desired characteristics. According to Mooney and Duval (1993), Bootstrapping is a type of database simulation that involves resampling data many times to produce empirical estimates of the set entire statistical sampling distribution. The basic concept of Bootstrapping is to use available data to estimate the error obtained from the empirical distribution function of the samples generated (Davidson & MacKinnon, 1993). Therefore, in research related to the stability of item parameter estimation, it is necessary to retrieve data using the Bootstrapping method.

(7) Estimating item and ability parameter on boostrap sample data. (8) The parameter estimation results produce an RMSD value in each replication with bootstrap data samples. In research conducted by Şahin and Anıl (2017), one of the accuracy criteria of grain parameter estimation uses Root Mean Squared Difference (RMSD). RMSD is defined by Equation (10), where $\bar{\pi}_i$ represents one of the estimated item parameters (a, b, or c) on item i, $\pi_i$ represents the corresponding item parameter baseline, taken as the true item parameter for item 1 and n is the number of items (Swaminathan et al., 2003).

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} (\pi_i - \bar{\pi}_i)^2}{n}}, \quad \text{(10)}$$

## FINDINGS AND DISCUSSION

### Findings

Data processing was done using Rstudio software. The language proficiency test data is estimated using the Rasch, 1PL, 2PL, and 3PL models and using the Items Characteristic Curve (ICC) graph. The results obtained from the three models will be compared based on the number of items that match. After getting the best model, the assumptions of unidimensionality, parameter invariance, and local independence are tested. After the three assumptions are met, Bootstrapping is performed on the original data to take the desired number of participants (100, 150, 250, 300, 500, 700, 1,000, 1,500, 2,000, 2,500) and replicated ten times to test the stability of item and ability parameter estimates. Ten samples with ten replications that have been made are used to estimate item and ability parameters using the best model on the original data. After obtaining the discriminant value (a), the difficulty parameter (b), and if using the 3PL model, an additional

pseudo guess parameter (c) is required. From the estimation results on each sample, the RMSD value will then be calculated. Parameter estimation is said to be stable if, from a sample, ten replications have a small RMSD value and are close to each other.

***Model Fit Test***

The best model selection can be made by testing the model fit with Yen's Q1 method and using the ICC graph. The value of $Q_1$ will be compared with $X^2$ table with degree of freedom $10 \times (J - 1) - m$. Thus, the Rasch model on dichotomous data has a free degree of 9. An item is said suitable if $Q_1 < X^2_{0,05(9)} = 16{,}92$ in the Rasch model fit test and 1PL which only estimates one parameter, namely the difficulty parameter (b), $Q_1 < X^2_{0,05(9)} = 15{,}51$ in the 2PL model fit test since it estimates two parameters, namely the difference parameter (a) and level of difficulty (b), and for the 3PL model fit test involving the difference parameter (a), the level of difficulty (b), and the pseudo guess (c), that is if $Q_1 < X^2_{0,05(7)} = 14{,}07$. Based on data analysis using the Rasch model, 1PL model, 2PL model, and 3PL model, a summary of the model fit test is shown in Table 1.

Table 1. Summary of Model Fit Test Results on the Four Models

| Decision | Rasch Model | 1PL Model | 2PL Model | 3PL Model |
|---|---|---|---|---|
| Suitable | 11 | 8 | 15 | 13 |
| Not Suitable | 35 | 38 | 31 | 33 |

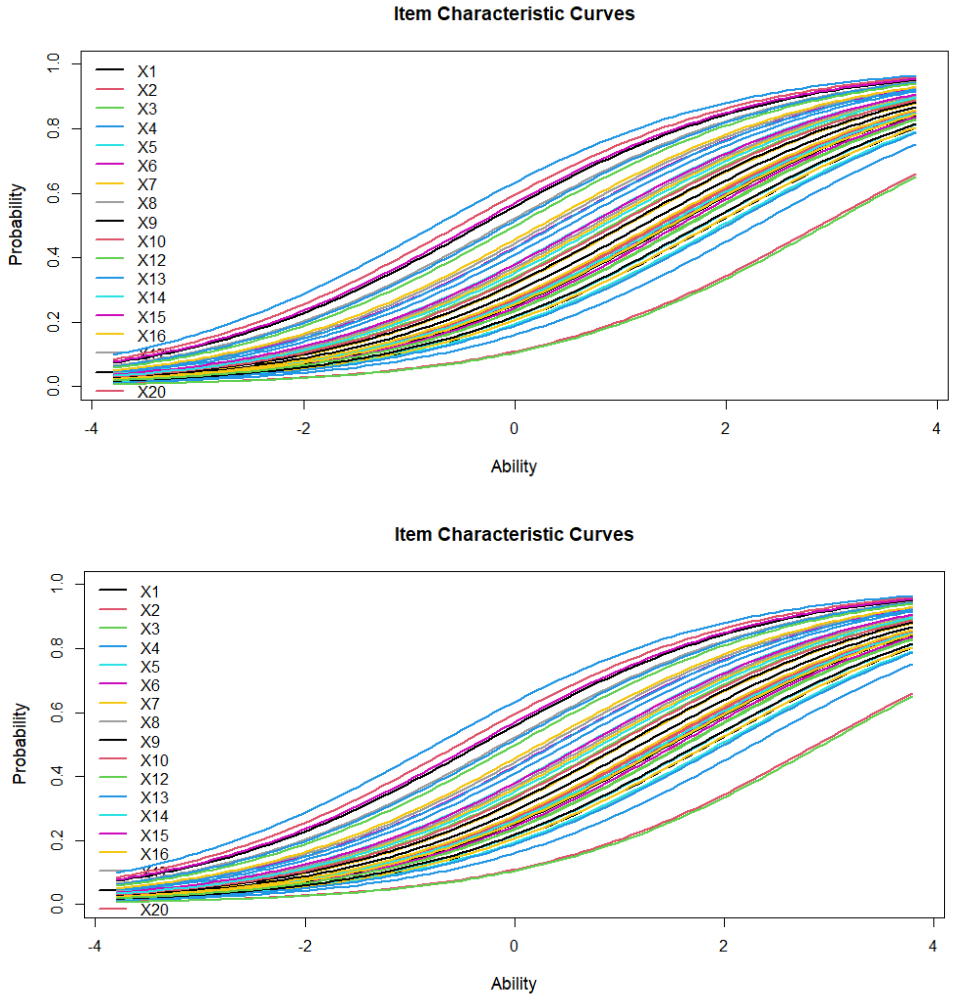


Figure 1. ICC Plot of Rasch and 1PL Models

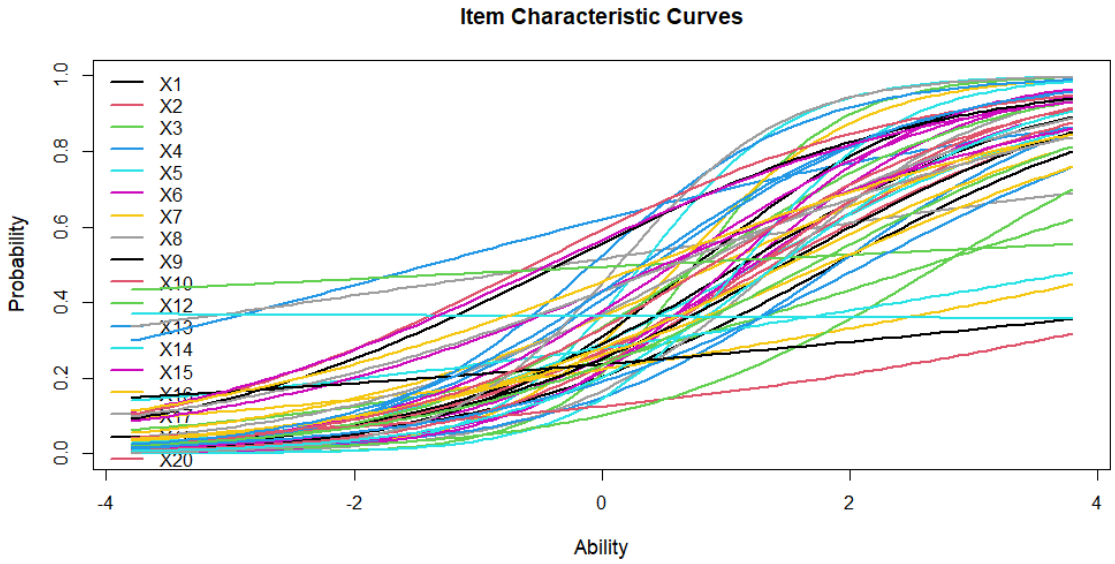


Figure 2. ICC Plot of 2PL Model

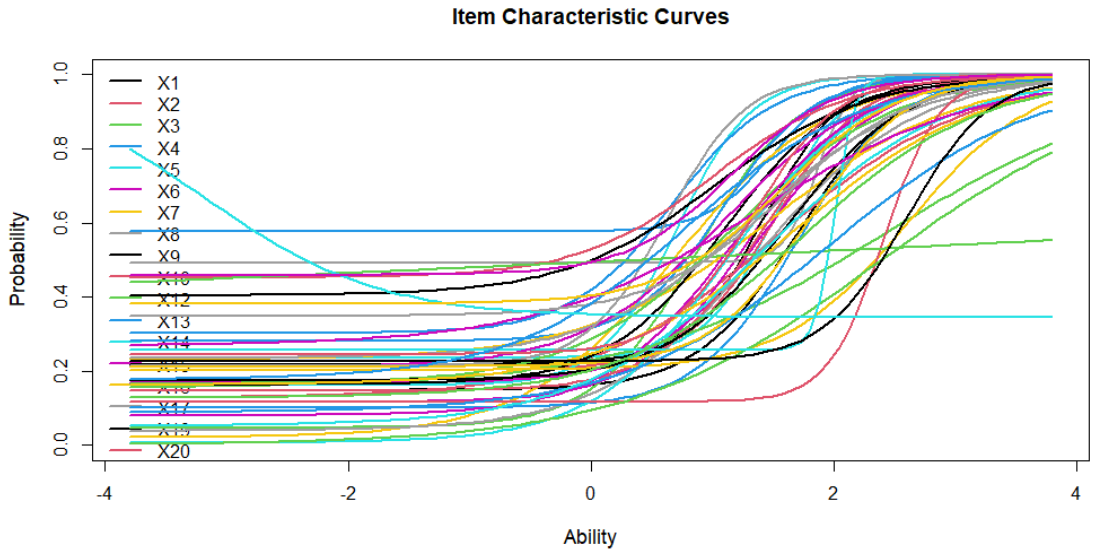


Figure 3. ICC Plot of 3PL Model

The ICC graph can be used to select the best model. The ICC plots for the Rasch and 1PL models are presented in Figure 1, while the 2PL model is in Figure 2, and the 3PL model is in Figure 3.

From Table 1, it can be seen that the 2PL model has the highest number of matching items. Thus, it can be concluded that, using the model fit test, the 2PL model is the best model for English listening test data, because it has the most number of items that match. By looking at the graph on the ICC plot of each model, it can be concluded that the Rasch model and the 1PL model as in Figure 1 are the best plots because the resulting graph forms an S curve. In Figure 2, the ICC plot of the 2PL model although some items produce graphs that do not follow the normal ogive and some do not form an S curve but are still acceptable, whereas in Figure 3, the 3PL ICC plot produces a graph where most of the items do not follow the normal ogive and do not form an S curve. Therefore, both using the model fit test and the 2PL model graph can be

selected as the best model because it has the highest number of suitable items and the resulting ICC plot is also quite good. In this study, the 2PL model was used as the best model for conducting research related to the stability of item parameter estimation considering the number of test participants on the data of the test device measuring English listening skills.

***Assumptions of Item Response Theory***

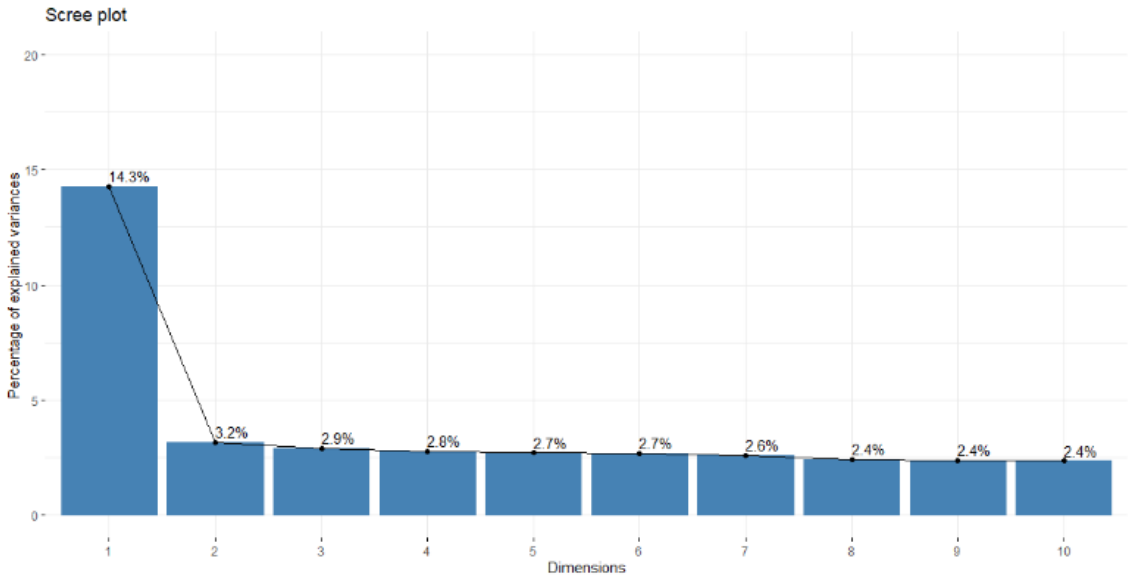


Figure 4. Scree Plot of the Data

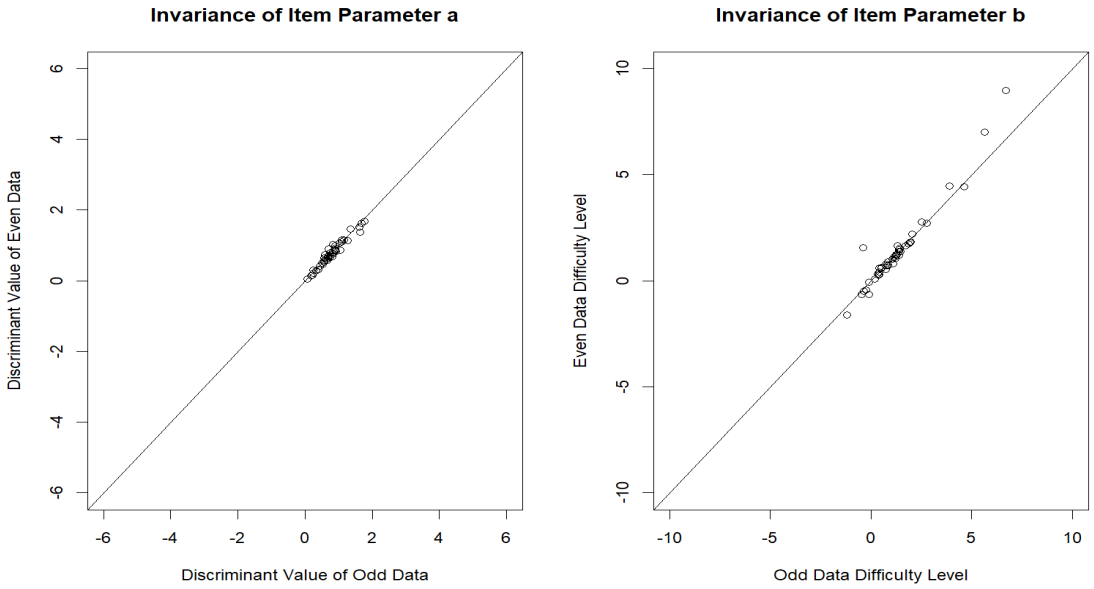


Figure 5. Parameter Invariance of Differential Item Functioning (a) and Difficulty Level (b)

The first assumption is unidimensionality. In Figure 4, it can be seen that dimension one is dominant over the other dimensions. In dimension one to dimension two there is also an elbow point, so it can be concluded that the unidimensional assumption is fulfilled. Because the assumption of unidimensionality has been met, the assumption of local independence has also been met (Hambleton et al., 1991). The local independence test is fulfilled if the answer between participants to an item does not affect the participants's answer to another item.

The next assumption is parameter invariance. In straight Figure 5, it can be seen that the points spread around the line and follow a line. Thus, it can be concluded that the assumption of invariance of item parameters of differentiation (a) and difficulty level (b) on the data of the test device measuring English/listening skills is fulfilled. Meanwhile, in Figure 6, it can be seen that the points also spread around the line and follow a straight line. Thus, it can be concluded that the assumption of invariance of ability parameters on the data of the language ability measurement test device is fulfilled.

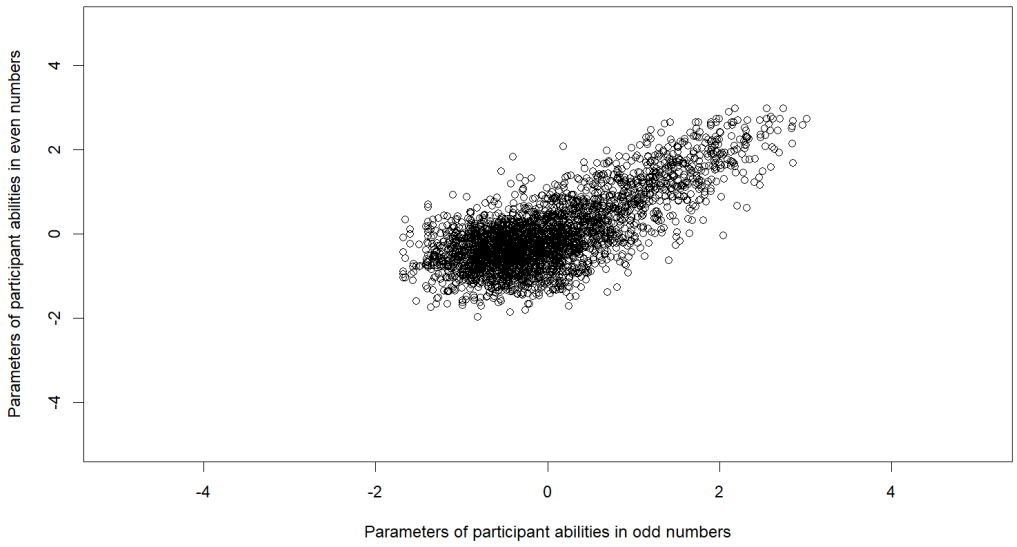


Figure 6. Invariance of Ability Parameters

***Effect of the Number of Test Takers on the Stability of Item Parameter Estimates***

Estimation of item parameters using the best model, namely 2PL on the original data, will get the estimated value of the discriminant value or differential power (a) and the difficulty level parameter (b). Criteria for item goodness according to Hambleton et al. (1991), namely, if the (a) is in the range $0 \leq a \leq 2$ and for the difficulty parameter (b) if it is in the range $-2 \leq b \leq 2$. From the analysis of item parameters with the 2PL model conducted, the results of estimating the discriminant or discriminating value (a) and the difficulty parameter (b).

The results of the discriminant value or power difference (a) in the original data obtained are then estimated using Equation (10) with the discriminant value or power difference (a) in the bootstrap sample data to obtain the RMSD value. Ten samples with ten replications were estimated and a summary of the ten RMSD values on the discriminant value or power difference (a) was obtained as in Table 2.

Table 2. Summary of RMSD Values on Discriminant Value or Discriminating Power (a)

| Replication | n100 | n150 | n250 | n300 | n500 | n700 | n1000 | n1500 | n2000 | n2500 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.378 | 0.263 | 0.193 | 0.165 | 0.143 | 0.111 | 0.096 | 0.071 | 0.059 | 0.077 |
| 2 | 0.354 | 0.293 | 0.188 | 0.199 | 0.143 | 0.101 | 0.134 | 0.068 | 0.061 | 0.067 |
| 3 | 0.336 | 0.247 | 0.222 | 0.168 | 0.149 | 0.116 | 0.092 | 0.068 | 0.086 | 0.056 |
| 4 | 0.238 | 0.221 | 0.202 | 0.168 | 0.137 | 0.142 | 0.079 | 0.070 | 0.070 | 0.076 |
| 5 | 0.261 | 0.279 | 0.264 | 0.156 | 0.148 | 0.105 | 0.085 | 0.080 | 0.074 | 0.053 |
| 6 | 0.411 | 0.218 | 0.201 | 0.186 | 0.130 | 0.095 | 0.099 | 0.090 | 0.067 | 0.080 |
| 7 | 0.318 | 0.276 | 0.171 | 0.178 | 0.104 | 0.106 | 0.094 | 0.081 | 0.073 | 0.045 |
| 8 | 0.360 | 0.283 | 0.180 | 0.182 | 0.128 | 0.125 | 0.084 | 0.075 | 0.076 | 0.062 |
| 9 | 0.422 | 0.215 | 0.236 | 0.170 | 0.182 | 0.115 | 0.093 | 0.067 | 0.070 | 0.065 |
| 10 | 0.282 | 0.302 | 0.183 | 0.194 | 0.143 | 0.113 | 0.100 | 0.082 | 0.078 | 0.068 |

When viewed from the summary table of RMSD values as in Table 2, it can be seen that the larger the sample or the greater the number of participants, the RMSD value on the discriminant value or differential power (a) is smaller, which can be interpreted as better. In the Bootstrap data sample of 100, it can be seen that the RMSD value generated between replications is not yet stable. Then, the data with a sample of 700 already shows stability between replications. For more details, a plot of the RMSD value for each bootstrap sample data with ten replications will be presented to facilitate drawing conclusions, as in Figure 7.

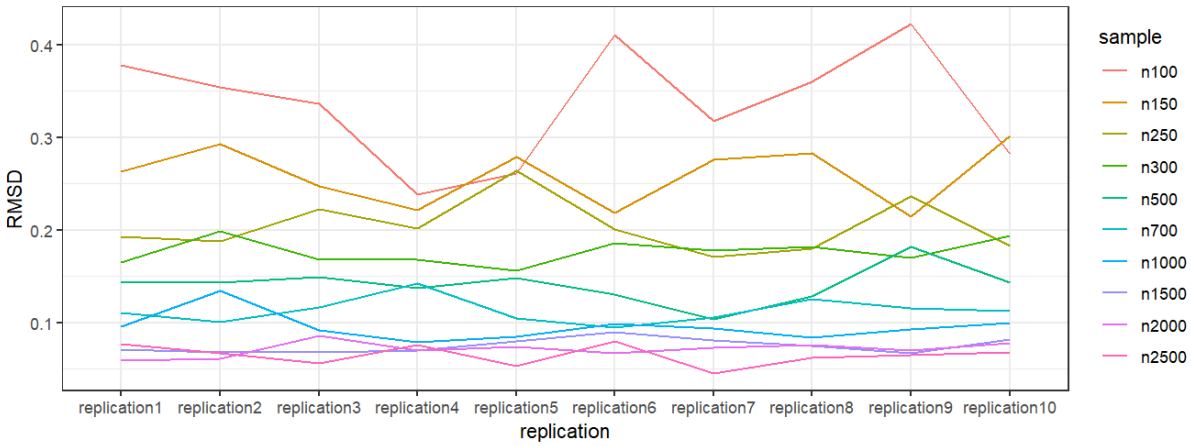


Figure 7. Plot of RMSD Value of Discriminant Value (a)

In Figure 7, it can be seen that the sample data of 700 RMSD values generated from The tenth replication has stabilized slightly, as seen from the shape of the line plot, which is close to a straight line. Figure 7 also shows that the more samples used with ten replications, the more stable the resulting RMSD value. This is evident in the plot with 2,500 sample data, the resulting line plot forms a straight line. Compared to the 100 sample data, the 2,500 sample data plot looks very stable compared to the 100 sample data.

On the other hand, the results of the difficulty parameter value (b) in the original data obtained were then estimated using equation (10) with the difficulty parameter value (b) in the Bootstrap sample data to obtain the RMSD value. Ten samples with ten replications were estimated, and a summary of the ten RMSD values on the discriminant value or differential power (b) was obtained as in Table 3.

Table 3. Summary of RMSD Values on the Difficulty Level Parameter (b)

| Replication | n100 | n150 | n250 | n300 | n500 | n700 | n1000 | n1500 | n2000 | n2500 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.477 | 11.003 | 1.143 | 0.828 | 3.557 | 0.590 | 3.417 | 1.325 | 0.410 | 0.217 |
| 2 | 2.066 | 11.487 | 2.859 | 0.543 | 0.359 | 1.331 | 0.613 | 0.690 | 0.590 | 0.211 |
| 3 | 0.753 | 4.742 | 58.538 | 0.822 | 0.449 | 0.914 | 0.620 | 0.233 | 0.338 | 0.360 |
| 4 | 74.432 | 0.876 | 0.815 | 0.719 | 0.969 | 0.843 | 0.368 | 0.596 | 0.665 | 0.396 |
| 5 | 3.623 | 4.203 | 1.002 | 5.160 | 9.708 | 3.154 | 0.344 | 2.772 | 1.039 | 0.588 |
| 6 | 21.947 | 4.363 | 6.716 | 0.638 | 26.810 | 0.632 | 0.306 | 0.954 | 0.622 | 0.602 |
| 7 | 93.582 | 16.060 | 0.817 | 15.740 | 0.685 | 0.694 | 0.672 | 0.410 | 0.762 | 0.247 |
| 8 | 13.539 | 60.713 | 1.112 | 4.019 | 7.637 | 1.229 | 0.353 | 0.511 | 0.517 | 0.203 |
| 9 | 1.425 | 1.137 | 1.566 | 1.699 | 0.850 | 2.977 | 0.708 | 0.557 | 0.520 | 0.550 |
| 10 | 1.435 | 0.837 | 0.524 | 0.735 | 6.743 | 0.357 | 0.831 | 6.381 | 0.303 | 0.408 |

When compared to the RMSD value of the discriminant value (a), the RMSD value of the difficulty parameter only begins to look stable in sample data 1,000. To facilitate drawing conclusions, the RMSD value of the difficulty parameter (b) will be presented in the form of a plot for each bootstrap sample data, as in Figure 8.

The plot in Figure 8 shows that in the 700 sample data, the RMSD value generated from the ten replications is quite stable, as seen from the shape of the line plot, which is close to a straight line. Moreover, it can also be seen that the more samples used with ten replications, the more stable the resulting RMSD value. This is evident in the plot with 2,500 sample data, the resulting line plot forms a straight line. Compared to the 100 sample data, the 2,500 sample data plot looks very stable compared to the 100 sample data.

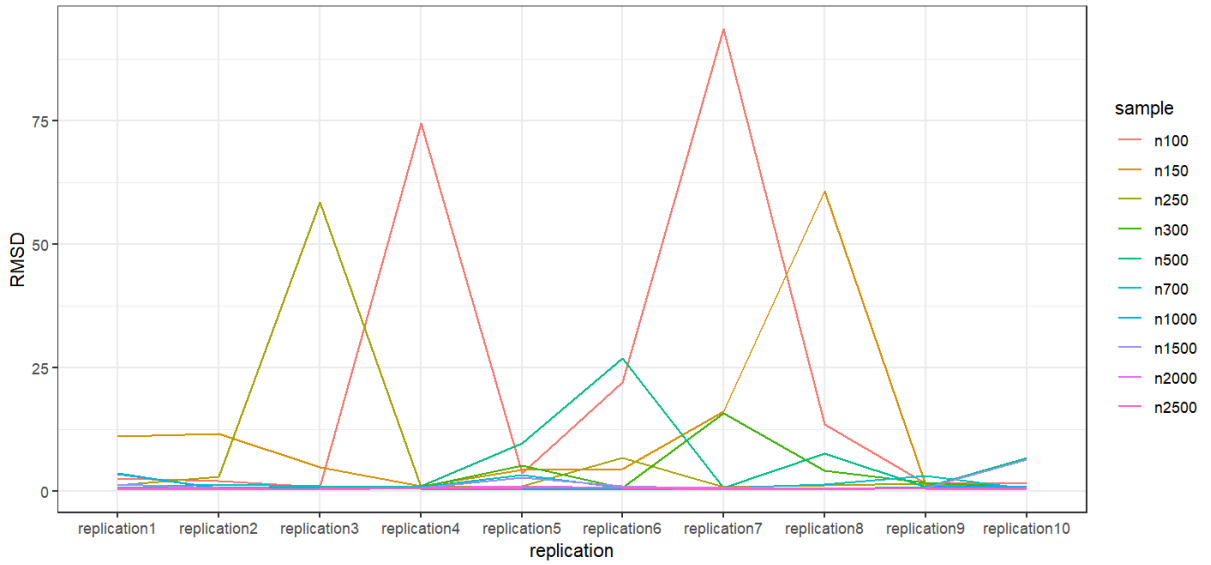


Figure 8. Plot of RMSD values of difficulty level parameter (b)

## Discussion

This study aims to find the stability of item parameter estimation by considering the number of test takers on dichotomous data using item response theory on the results of the test device measuring English listening skills using the Bayes method (expected a posteriori, EAP). Based on the model fit test as in Table 1 and using the ICC graph method, the best or most suitable model for parameter estimation on English listening test data is the 2PL model. The number of items that fit in this 2PL model is 15 items, which is the model with the most suitable items. The graph on the ICC plot produced by the 2PL model is also included in the good and acceptable category because it follows the normal ogive even though several items do not follow the normal ogive shape. Thus, it can be concluded that the 2PL model was chosen to be the best model.

In the results of the RMSD value on the discriminant value (a), as in Table 2, it can be seen that the RMSD value on the 700 sample data with ten replications has begun to appear stabilized. Based on the plot of the results of the RMSD value of the discriminant value (a) of the Bootstrapping sample data, as shown in Figure 7, the sample data of 100 is relatively large and unstable. For sample data, 150 also looks unstable, but the resulting RMSD value is smaller than that of the data in sample 100. In the 250 and 300 sample data, the RMSD value becomes smaller and stabilises. Then, in the 500 sample data, the RMSD value is considered stable. For the next sample data, namely 700, 1,000, 1,500, 2,000, and 2,500, the resulting RMSD value is getting smaller and more stable. The more data in the sample, the more the resulting line plot forms a straight line, which means stability is met. Therefore, as the results obtained in Table 2, the

number of test participants affects the estimation of item parameters. The more sample data of test participants, the smaller the RMSD value and the more stable the plot. This indicates that the larger the sample, the better the item parameter estimation.

Likewise, in the results of the RMSD value of the difficulty level parameter (b), as shown in Figure 8, the 100 sample data is very large and unstable. For sample data 150, it also looks unstable. The RMSD values in Table 3 generated in sample data 100, 150, 250, 300, and 500 are very large and unstable in all ten replications. In the 700 sample data, the RMSD value with ten replications produced is not as large as the previous five sample data and looks more stable. However, the RMSD value of the difficulty level parameter (b) in the 700 sample data is still relatively large and not so good. In the next sample data, namely 1,000, 1,500, 2,000, and 2,500, the resulting value is more stable. The RMSD value of the difficulty level parameter (b) decreased and stabilized in the sample data of 1,000 participants. Thus, in this study, it can be concluded that the stability of item parameter estimation with the Bayes method (Expected A posteriori, EAP) using the 2PL model is with 1,000 test participants. Research related to the stability of item parameters has been conducted. It was found that the number of test participants affects the stability of item parameter estimates in the 2PL IRT model. This result is supported by research conducted by Stone and Yumoto (2004), Custer (2015), and Akour and Al-Omari (2013) with the same method but applied to different models and data, that the sample size affects the stability of item parameter estimates.

## CONCLUSION

Based on the results of research related to the stability of item parameter estimation and ability on dichotomous data of test devices measuring English listening skills in 2021, it can be concluded that the most suitable model used to estimate this data in item response theory is the 2PL model. The 2PL model is used to find the stability of the item parameter estimate and the result is that the number of test participants is a variable that can affect the estimation results. The results obtained for the minimum data sample on the stability of the item parameter estimate are on test participants with a total of 1,000 participants. Suggestions that can be given for further research are to test the stability of item parameter estimates by considering other variables that are thought to influence such as test length or a combination of many test takers and test length. Further research is recommended to use other methods in IRT, for example, by using the Maximum Likelihood Estimation (MLE) method or the Maximum A Posteriori (MAP) method to compare the results obtained. Because of the results obtained in this study that the minimum data sample on the stability of item parameter estimation is on test participants with a total of 1,000 participants, it is recommended for researchers who want to estimate item parameters to use more than 1,000 test participants.

## DISCLOSURE STATEMENT

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

## REFERENCE

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. https://journals.indexcopernicus.com/search/article?articleId=608863

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

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

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

Crivelli, D., Spinosa, C., Angelillo, M. T., & Balconi, M. (2023). The influence of language comprehension proficiency on assessment of global cognitive impairment following Acquired Brain Injury: A comparison between MMSE, MoCA and CASP batteries. *Applied Neuropsychology: Adult, 30*(5), 546–551. https://doi.org/10.1080/23279095.2021.1966430

Custer, M. (2015). Sample size and item parameter estimation precision when utilizing the one-parameter "Rasch" model. Paper presented at the *Annual Meeting of the Mid-Western Educational Research Association Evanston*, Illinois October 21-24.

Davidson, R. & MacKinnon, J.G. (1993). *Estimation and inference in econometrics*. Oxford University Press.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press Taylor & Francis Group.

Falani, I., Nisraeni, N., & Irdiyansyah, I. (2018). The ability of estimation stability and item parameter characteristics reviewed by Item Response Theory model. *Proceedings of the International Conference on Education in Muslim Society (ICEMS 2017)*. Atlantis Press. https://doi.org/10.2991/icems-17.2018.34

Gong, Y. (Frank), Chen, M., & An, Z. (2024). Examining the impact of foreign language anxiety and language contact on oral proficiency: a study of Chinese as a second language learners. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2023-0328

Guo, S., Wu, T., Zheng, C., & Chen, Y. (2021). Bayesian modal estimation for the one-parameter logistic ability-based guessing (1PL-AG) model. *Applied Psychological Measurement, 45*(3), 195–213. https://doi.org/10.1177/0146621621990761

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). Macmillan Publishing Co, Inc.; American Council on Education.

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

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

Kassambara, A., & Mundt, F. (2020). *Package 'factoextra': Extract and visualize the results of multivariate data analyses*. CRAN- R Package, 84. https://cran.r-project.org/package=factoextra

Khodi, A., Ponniah, L. S., Farrokhi, A. H., & Sadeghi, F. (2024). Test review of Iranian English language proficiency test: MSRT test. *Language Testing in Asia, 14*(1), 4. https://doi.org/10.1186/s40468-023-00270-0

Kim, A. A., Yumsek, M., Kemp, J. A., Chapman, M., & Gary Cook, H. (2023). Universal tools activation in English language proficiency assessments: A comparison of Grades 1–12 English learners with and without disabilities. *Language Testing, 40*(4), 877–903. https://doi.org/10.1177/02655322221149009

Mooney, C. Z. & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage Publication.

Neiriz, R. (2023). *Eliciting interactive oral communication samples through a spoken dialogue system to measure interactional competence*. Thesis, Iowa State University. https://doi.org/10.31274/td-20240617-300

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

Sabitova, K. (2023). Determination and assessment of students' basic competence level in school education. *Общество и Инновации, 4*(10/S), 219–226. https://doi.org/10.47689/2181-1415-vol4-iss10/S-pp219-226

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Kuram ve Uygulamada Egitim Bilimleri*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Sarac, M., & Loken, E. (2023). Examining patterns of omitted responses in a large-scale English language proficiency test. *International Journal of Testing, 23*(1), 56–72. https://doi.org/10.1080/15305058.2022.2070756

Schleicher, I., Leitner, K., Juenger, J., Moeltner, A., Ruesseler, M., Bender, B., … Kreuder, J. G. (2017). Examiner effect on the objective structured clinical exam – a study at five medical schools. *BMC Medical Education, 17*(1), 71. https://doi.org/10.1186/s12909-017-0908-1

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating Rasch/IRT parameters with dichotomous items. *Journal of Applied Measurement*, *5*(1), 48–61.

Susongko, P. (2021). The estimation stability comparison of participants' abilities on scientific literacy test using Rasch and One-Parameter Logistic model. *Journal of Physics: Conference Series, 1842*(1), 012037. https://doi.org/10.1088/1742-6596/1842/1/012037

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, 27(1), 27–51. http://dx.doi.org/10.1177/0146621602239475

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262. https://doi.org/10.1177/014662168100500212