BAREKENG: Journal of Mathematics and Its Applications
December 2026 Volume 20 Issue 4 Page 3259-3270
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss4pp3259-3270

MULTILEVEL ITEM RESPONSE THEORY MODEL USING MML-GHQ METHOD FOR HIERARCHICAL DATA

Alona Dwinata[1*], Anang Kurnia[2], Aji Hamim Wigena[3], Muhammad Nur Aidi[4]

[1,2,3,4] *School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Raya Dramaga, Bogor, 16680, Indonesia

*Corresponding author's e-mail: *alonadwinata@umrah.ac.id*

Article Info

Article History:

Received: 8th October 2025
Revised: 27th March 2026
Accepted: 31st May 2026
Available online: 24th August 2026

Keywords:

Gauss-Hermite Quadrature;
Hierarchical Data;
Maximum Marginal Likelihood;
Multilevel Item Response Theory.

ABSTRACT

Hierarchical item response data require a multilevel approach to capture the diversity of respondent abilities and item characteristics by separating variations between examinees and schools to make item parameter and ability estimates more precise. This study aims to develop a two-parameter logistic multilevel item response theory (MIRT 2PL) model, employing the maximum marginal likelihood method with Gauss-Hermite quadrature (MML-GHQ) to estimate the rank order of examinees' abilities. This study specifically investigates the efficiency and accuracy of the MIRT 2PL model with MML-GHQ to predict the ability rankings of examinees. The research incorporates both simulated and empirical data. The simulation study generated item response data under a two-level hierarchical structure, where examinees were nested within schools. The population consisted of 50 schools, with 20–30 students per school and five items. Each examinee's ability was modelled as a combination of school-level and individual-level effects, under two conditions of school variability: high ($\tau = 1.2$) and low ($\tau = 0.6$). Random samples were drawn from the population, and the sampling process was repeated 10 times to assess consistency. The analysis included estimating item parameters, variance components, and examinees' abilities using the EAP approach. Model performance was evaluated using RMSE, Spearman correlation, and computation time. Results indicated that MML-GHQ produced accurate and consistent rank estimates, particularly under high school variability. Using PISA data, increasing the number of schools, students, and items in the empirical data yielded results consistent with the simulation study. In conclusion, the MIRT 2PL model with MML-GHQ offers an effective and efficient alternative for estimating ability rankings in hierarchical item response data.

How to cite this article:

A. Dwinata, A. Kurnia, A. H. Wigena, and M. N. Aidi, "MULTILEVEL ITEM RESPONSE THEORY MODEL USING MML-GHQ METHOD FOR HIERARCHICAL DATA", *BAREKENG: J. Math. & App.*, vol. 20, no. 4, pp. 3257-3268, Dec. 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Measurement in education requires an instrument in the form of a collection of test items used to measure test participants' abilities. Data on test participants' answers to each test item is called item response data. The analysis of item response data based on its objectives uses two approaches, namely, the measurement approach and the explanatory approach. The measurement approach aims to determine the abilities of test participants. Meanwhile, the explanatory approach analyzes predictors that significantly influence test participants' abilities. Item response data with a categorical measurement scale is analyzed using the Item Response Theory (IRT) model [1]. The measurement of test participants' abilities in the IRT model is unidimensional, meaning that only one ability underlies each test participant's answer [2]. The IRT model explains the probability of test participants' answers on a test item as a function of the test participants' ability characteristics and the characteristics of the test item [3]. The IRT model requires two stages of analysis to complete the analysis with an explanatory approach, namely the measurement approach and the explanatory approach, because predictors cannot be used directly in the model. The IRT model's limitations can be overcome using the Generalized Linear Mixed Models (GLMM) approach [4]. Previous research has also formulated an IRT model from a Hierarchical Generalized Linear Model (HGLM) perspective, with item response data (level 1) nested within examinees (level 2) [5]–[7]. The hierarchical data structure arises from the population condition where sample units can be grouped at higher levels.

Hierarchically structured data does not meet the assumption of independence between observations. Previous research has shown that ignoring hierarchical structure in data can lead to biased parameter estimates [8]. The same applies to hierarchically structured item response data; ignoring the hierarchical structure in the model has negative consequences for the estimated standard error [9]. The hierarchical structure in the data is accounted for using a multilevel model [10]. The IRT model, combined into a multilevel model structure, is called the multilevel IRT (MIRT) model. MIRT measures examinees' ability by explaining differences at different ability levels. Response variables are at level 1, and explanatory predictors can be added at all levels [11]. MIRT modeling with examinee predictors on purely hierarchically structured data is modeled from a generalized multilevel perspective as a three-level multilevel model [12]. The three-level MIRT 1PL model in educational data testing examines examinees and item parameters and local dependencies between examinees [13]. The MIRT 1PL model assumes all items have equal discrimination power. One study found that differences in item discrimination power across levels influence the correlation between individual and group abilities [14], [15].

Previous researchers have researched the MIRT 2PL model. Examinees' ability parameters in the MIRT 2PL model can be estimated using the PLmixed package in RStudio [16]. Previous research has considered the hierarchical structure of item response data, but there has been no study on estimating examinees' ability rankings. Studies on examinees' ability rankings are essential during the selection and evaluation process in education. Researchers are interested in examining the stability and consistency of examinees' ability rankings using the MIRT 2PL model with the Maximum Marginal Likelihood (MML) method. The marginal likelihood function is approximated using the Gauss-Hermite Quadrature (GHQ) numerical method [17]. Researchers proposed the Maximum Marginal Likelihood method with the Gauss-Hermite Quadrature numerical method, or MML-GHQ, to see the consistency of the estimated examinees' ability ranking order compared to the actual ranking order.

The real problem of this research is modeling item response data in the Programme for International Student Assessment (PISA). The PISA participant population is children aged 15 years who are currently studying. PISA sampling is carried out in two stages. The first stage is carried out by randomly selecting schools, and in the second stage, examinees are randomly selected in each sample school. Consequently, examinees are nested within schools, creating a hierarchical data structure that may influence response patterns and the distribution of abilities across groups. This sampling process requires a multilevel model to separate variation in ability between exams and schools because school factors (e.g., resources, learning environment, or policies) can systematically influence students' achievement. Therefore, accounting for this structure is necessary to obtain efficient and accurate estimation under the MIRT 2PL model with MML-GHQ and to ensure that the resulting ability rankings are reliable. Examinees' ability estimates become more precise by including the random effect of school because they account for the influence of different learning environments across schools. This study analyzes explicitly item response data in the mathematics domain, where the ability measured is mathematical literacy. This study aims to estimate the ranking order of students' mathematical literacy abilities using the MIRT 2PL model with the MML-GHQ method. By including the random effect of school, the ability estimates become more precise because they account for the influence of different learning environments across schools. The model can show how much of the variance in ability stems from differences between schools. The MIRT 2PL model allows each item to have its own discriminant and difficulty level, making the analysis more realistic and sensitive to differences in PISA item characteristics. It can also identify quality gaps between schools in education policy analysis.

2. RESEARCH METHODS

Based on the study's background and objectives, this research began with the development of the MIRT 2PL model using the MML-GHQ method. Next, the researcher conducted simulations and empirical studies on the developed model. The tool used in this study was RStudio.

2.1 Development of the Two-Parameter Logistic MIRT Model

The two-Parameter Logistic IRT (IRT 2PL) model has two item parameters: the item difficulty with a value between 0.00 and 2.00, and the item discrimination with a value between -3.00 and +3.00. The IRT 2PL model is formulated in the form of an intercept (α_i) and an item loading (λ_i), which are unique [1], [3]. The IRT 2PL model is written as Eq. (1).

$$P\left(y_{ij}=1 \middle| \theta_j, \alpha_i, \lambda_i\right)=\frac{e^{\alpha_i+\lambda_i\theta_j}}{1+e^{\alpha_i+\lambda_i\theta_j}}. \tag{1}$$

With:

θ_j: the ability of the j-th examinee;

λ_i: item discrimination parameters and $\frac{-\alpha_i}{\lambda_i}$ = the difficulty of test items.

The IRT 2PL model for item response data with examinee ability dependent on the random effect of school is called the MIRT 2PL model.

$$\operatorname{logit}\left[P\left(y_{ijk}=1\right)\right]=\alpha_i+\lambda_i\theta_{jk},$$
$$\theta_{jk}=\delta_{jk}+\delta_k, \text{with } \delta_{jk}\sim N(0,\sigma^2) \text{ and } \delta_k\sim N(\mu_\delta,\tau^2),$$

With:

θ_{jk}: a random effect of the ability of the j-th examinee in the k-th school;

δ_{jk}: a random effect of students within a school;

δ_k : a random effect of the school;

$P\left(y_{ijk}=1\right)$: the probability that the j-th examinee at the k-th school answers correctly on the i-th item

MIRT 2PL in the GLMM approach is stated as Eq. (2).

$$\operatorname{logit}\left[P\left(y_{ijk}=1\right)\right]=\alpha_i+\lambda_i\left(\delta_{jk}+\delta_k\right). \tag{2}$$

2.2 Simulation Study

The Simulation data were generated based on two-level hierarchical item response data, where test participants are nested within schools. This study used two scenarios: school population data with high and low levels of diversity. Item response data population with high school diversity ($\tau = 1.2$) and low school diversity ($\tau = 0.6$). The procedure for generating item response data with high school diversity ($\tau = 1.2$) is as follows:

1. Population data generation
   a. Generating population data from 50 schools, with the number of test participants varying randomly between 20 and 30 test participants for each school, each test participant answers five questions.
   b. Determine the actual population parameter values, namely $\mu_\delta = 0.5$ (mean school effect), $\tau = 1.2$ (standard deviation of school effect), and $\sigma = 0.3$ (standard deviation of test participant effect)

c. Generate the school effect (δ_k) using Normal distribution $(0.5, 1.2)$ and the examinee effect within schools (δ_{jk}) using Normal distribution $(0, 0.3)$. Calculate the test taker's ability (θ_{jk}) using $\theta_{jk} = \delta_{jk} + \delta_k$.

d. Determine the parameter values for 5 test items, with the first item serving as a reference item with parameter values $\alpha = 0$ (item difficulty level) and $\lambda = 1$ (item discrimination power). The α parameters for the other four items were generated randomly from the Uniform distribution $[-2{,}2]$, and the λ parameters for the other four items were generated through exponential transformation of the Uniform distribution $[-0.7{,}0.8]$.

e. Generate response data for each examinee (y_{ijk}) by calculating the probability of correctly answering the i-th item using the 2PL IRT model

$$P(y_{ijk}) = \text{logit}^{-1}(\alpha_i + \lambda_i \theta_{jk}),$$

y_{ijk} was generated using the Bernoulli distribution.

f. Store population data in "Population" and "Examinee's Ability". The "Population" data contains the school's ID, examinee's ID, examinee's ability score (TrueTheta), and examinee's response to each question item. The "Examinee's Ability" data stores the examinee's ability scores (TrueTheta), school, and individual effects.

2. Sampling procedure

Take a random sample of 20 schools from the school population. All variables in the population data for each sample school are also stored in two data frames. Namely "Sample" and "Sample of Examinee's Ability".

3. Data analyst

a. Prepare the item response data structure for the sample data.

b. Estimate the parameters of the MIRT 2PL model using the MML-GHQ method.

c. Analyze the estimated parameters, namely the item parameters $\hat{a}_i$ and $\hat{b}_i$, the random effect diversity parameters σ^2 and τ^2.

d. Estimate the examinee's ability parameters using EAP.

e. Determine the RMSE value and Spearman correlation between the estimated examinee's ability parameter values and the actual examinee's ability value.

f. Observe the computation time of the MIRT 2PL model using the MML-GHQ method.

g. Repeat data analysis steps 1 through 6 by repeating the random sampling of schools 10 times to evaluate model performance.

To generate item response data under the low school diversity ($\tau = 0.6$), we repeated the same simulation procedure used for the high school diversity, changing only the school-level variance parameter to represent weaker between-school differences. Specifically, in the population data generation stage in step 2, we set $\tau = 0.6$, while keeping all other model parameters and simulation steps unchanged. This design ensures that differences between conditions are attributable solely to the magnitude of school-level heterogeneity.

2.3 Empirical Study

The empirical data analysis in this study aims to estimate students' mathematical literacy abilities using the MIRT 2PL model with the MML-GHQ method. The empirical data used in this study are the 2018 Indonesian PISA data. The PISA participant population was 85% of Indonesian children aged 15 [18]. Student sampling was done in two stages, so we had hierarchically structured item response data with students nested in schools. The total sample in this study was 1302 students from 381 schools. The items used were from the mathematics domain because the researcher focused on determining students' mathematical literacy abilities. The researcher determined that the item response data used were answers to 9 mathematics questions. The item response data in this study were binary with two possible answers for each item: a score of 1 if the student's answer was correct and a score of 0 if the student's answer was incorrect.

The stages of empirical data analysis for the MIRT 2PL model using the MML-GHQ method are as follows:

1. Prepare the item response data structure, item ID, student ID, and school ID in wide-form format for analysis using the MML-GHQ method.
2. Estimate the item parameters $\hat{a}_i, \hat{b}_i$, and students' mathematical literacy ability using the MML-GHQ method.
3. Analyze and interpret the estimated MIRT 2PL model parameters obtained using the MML-GHQ method.

3. RESULTS AND DISCUSSION

3.1 Development of the MIRT 2PL Model Using The MML-GHQ Method

The MIRT 2PL model is needed when the item response data is hierarchical. Eq. (2) states the IRT 2PL model as in Eq. (1) in the GLMM approach. The likelihood function of the item parameters and the random effect variance, assumed to be conditionally independent between responses. The likelihood function is stated as Eq. (3).

$$L(\alpha_i, \lambda_i, \sigma, \tau | y_{ijk}) = \prod_{k=1}^{K} \prod_{j=1}^{n_{jk}} \prod_{i=1}^{I} P(y_{ijk} | \theta_{jk}). \quad (3)$$

With:

K : number of schools;

n_{jk} : number of examinees in the k-th school;

I : number of items;

$P(y_{ijk}|\theta_{jk})$: the chance of an examinee's answer being correct depends on their ability.

The ability of the j-th examinee in the k-th school (θ_{jk}) cannot be observed directly and is modeled as a random variable. Parameter estimation using the MML method assumes that the test participant's ability is normal; errors in determining this distribution can result in bias [19], [20]. The MML approach integrates random effects into the MIRT 2PL model. The likelihood function is marginalized against the distribution of the random effects δ_{jk} and δ_k. The marginal likelihood function is written as Eq. (4).

$$L(\alpha_i, \lambda_i, \sigma, \tau | y_{ijk}) = \prod_{k=1}^{n_k} \left[\int_{-\infty}^{\infty} \prod_{j=1}^{n_{jk}} \left(\int_{-\infty}^{\infty} \left(\prod_{i=1}^{n_{ijk}} \frac{e^{y_{ijk}[\alpha_i + \lambda_i(\delta_{jk} + \delta_k)]}}{1 + e^{\alpha_i + \lambda_i(\delta_{jk} + \delta_k)}} \right) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\delta_{jk}^2}{2\sigma^2}} d\delta_{jk} \right) \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(\delta_k - \mu_\delta)^2}{2\tau^2}} d\delta_k \right], \quad (4)$$

with $\delta_{jk} \sim N(0, \sigma^2)$ and $\delta_k \sim N(\mu_\delta, \tau^2)$, The value of the log marginal likelihood is written as Eq. (5).

$$\log L(\alpha_i, \lambda_i, \sigma, \tau | y_{ijk}) = \sum_{k=1}^{K} \log \left[\int_{-\infty}^{\infty} \prod_{j=1}^{n_k} \left(\int_{-\infty}^{\infty} \left(\prod_{i=1}^{I} \frac{e^{y_{ijk}[\alpha_i + \lambda_i(\delta_{jk} + \delta_k)]}}{1 + e^{\alpha_i + \lambda_i(\delta_{jk} + \delta_k)}} \right) \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\delta_{jk}^2}{2\sigma^2}} d\delta_{jk} \right) \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(\delta_k - \mu_\delta)^2}{2\tau^2}} d\delta_k \right]. \quad (5)$$

With:

K : number of schools;

n_{jk} : number of examinees in the k-th school;

I : number of items;

λ_i : item discrimination parameters and $\frac{-\alpha_i}{\lambda_i}$ = the difficulty of test items

y_{ijk} : an examinee's answer (0 or 1);

δ_{jk}: a random effect of students within a school, $\delta_{jk} \sim N(0, \sigma^2)$;

δ_k : a random effect of the school, $\delta_k \sim N(\mu_\delta, \tau^2)$.

The integral of the marginal likelihood function is not in closed form; the integral value is approximated using the Gauss-Hermite Quadrature (GHQ) numerical method [17], as in Eq. (6).

$$\int_{-\infty}^{\infty} e^{-x^2} f(x)dx \approx \sum_{k=1}^{m} w_q f(x_q). \tag{6}$$

With:

m : number of quadrature points;

x_q : roots of the Hermite polynomial $H_m(x)$;

w_q : weight associated with a point x_q.

The model assumptions are structured at the examinee and school levels. The examinee's ability is $\theta_{jk} = \delta_k + \sigma\, z_{jk}$; $z_{jk} \sim N(0,1)$, and the school effect is modeled as $\delta_k = \mu_\delta + \sqrt{2}\tau v_k$; $v_k \sim N(0,1)$. After transforming the inner and outer integrals of the marginal likelihood function into the standard GHQ form, the likelihood values are obtained as in Eq. (7).

$$\log L \approx \sum_{k=1}^{K} \log\left(\frac{1}{\pi}\sum_{r=1}^{m} v_r \prod_{j=1}^{n_k}\left[\sum_{q=1}^{m} w_q \prod_{i=1}^{I} \frac{e^{y_{ijk}[\alpha_i+\lambda_i(\sqrt{2}\sigma x_q+\mu_\delta+\sqrt{2}\,\tau z_r)]}}{1+e^{\alpha_i+\lambda_i(\sqrt{2}\sigma x_q+\mu_\delta+\sqrt{2}\,\tau z_r)}}\right]\right). \tag{7}$$

Point z_r and weight v_r; $r = (1, \dots, R)$ for school-level integration.

Point x_q and weight w_q; $q = (1, \dots, Q)$ for examinee-level integration.

Quadrature point transformation as $\delta_k^{(r)} = \mu_\delta + \sqrt{2}\,\tau z_r$ and $\theta_{jk}^{(r)} = \sqrt{2}\sigma x_q + \mu_\delta + \sqrt{2}\,\tau z_r$

The Expected A Posteriori (EAP) approach estimates the examinees' ability parameters. The school effect is estimated using Eq. (8).

$$\hat{\delta}_k^{EAP} = \frac{\sum_{r=1}^{R}(\mu_\delta + \sqrt{2}\,\tau z_r)\cdot \exp\left(\log v_r + \sum_j \log P\left(u_j|\delta_k^{(r)}\right)\right)}{\sum_{r=1}^{R} \exp\left(\log v_r + \sum_j \log P\left(u_j|\delta_k^{(r)}\right)\right)}, \tag{8}$$

while the examinee's ability is estimated using Eq. (9).

$$\hat{\theta}_{jk}^{EAP} = \frac{\sum_{q=1}^{Q} \theta_{jk}^{(q)} \cdot \exp\left(\log P\left(u_j|\theta_{jk}^{(q)}\right)\right)\cdot w_q}{\sum_{q=1}^{Q} \exp\left(\log P\left(u_j|\theta_{jk}^{(q)}\right)\right)\cdot w_q}, \tag{9}$$

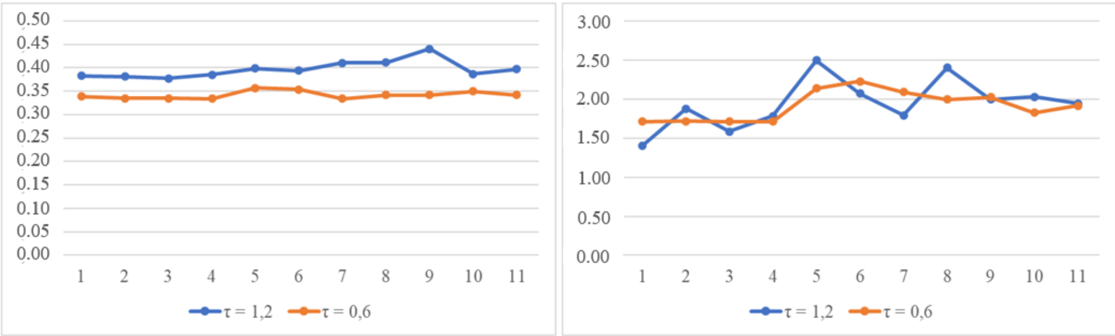
with

$$\theta_{jk}^{(q)} = \hat{\delta}_k + \sigma\sqrt{2}z_q.$$

Integration is done numerically with the number of points GHQ R and Q equal to 15.

3.2 Simulation Study

The MIRT 2PL model parameter estimation accuracy using the MML-GHQ method was analyzed based on the RMSE value, the Spearman correlation value between the actual ability and the estimated ability from the model, and the computation time. The results of the simulation data analysis with 10 sampling times from two different populations are shown in Fig. 1. The average RMSE of examinees' ability from a population with a school diversity parameter of 1.2 was 0.39, while the population with a school diversity parameter of 0.6 was 0.34.



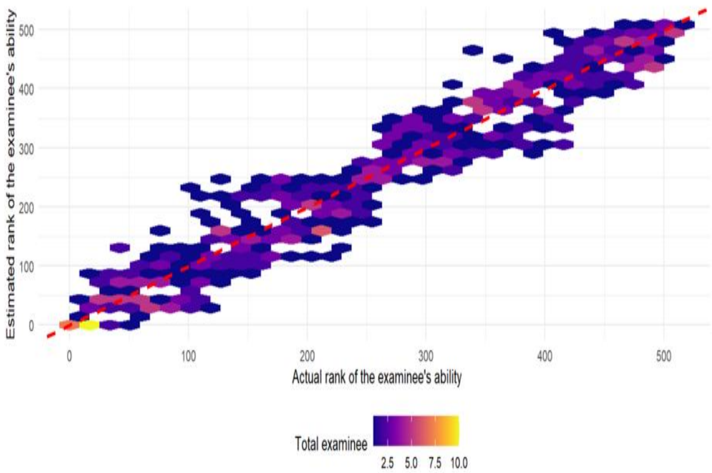
(a) (b)

Figure 1. The Results of The Simulation Data Analysis
(a) RMSE, (b) Computing Time

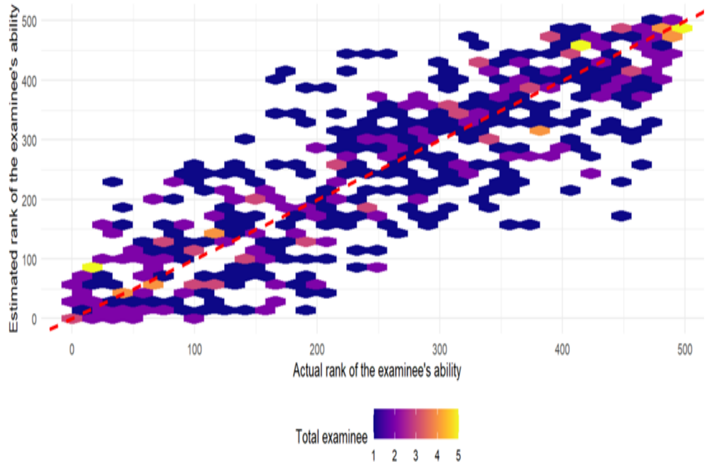
Fig. 1 (a) shows that the RMSE of examinees' abilities with high heterogeneous school effects has a higher RMSE value than that of low heterogeneous school effects. The RMSE value indicates that high diversity in the population results in the estimated ability of test participants being more dispersed, so that the RMSE value increases, resulting in the estimated ability of test participants being less precise. A lower school diversity value indicates that the influence of schools on examinees' abilities has a smaller source of diversity, so the estimated ability of the examinee is more accurate.

The performance of the MML-GHQ method is also assessed based on computational time. Fig. 1 (b) shows the computational time required to analyze simulated data with 10 samplings from two different populations. The analysis results indicate that the average difference in computational time for the two-sample data is only 1.69 seconds. Changes in the school's diversity value do not affect the computational burden in the parameter estimation process.

After determining the accuracy of the parameter estimates through the RMSE value and computational time efficiency, the researchers then analyzed the ability rankings to answer the objectives of this study. Spearman correlation measures the relationship between two estimated examinees' ability rankings and their actual abilities. Correlation analysis was performed on each estimated result of simulated data with 10 samples from two different populations. The data analysis results show that the average Spearman correlation value for examinees' abilities from the population with the parameter $\tau = 1.2$ was 0.94, while the population with parameter $\tau = 0.6$ was 0.82. A higher Spearman correlation value in the MIRT 2PL model with the MML-GHQ method indicates that the estimated order of examinees' abilities is more in line with the actual order of test participants' abilities. The graph of the suitability of the estimated rank of examinees' ability rankings for the two population data scenarios is shown in Fig. 2.



(a)



(b)

Figure 2. Graph Of Examinee Ability Rankings from Two Population Scenarios (a) School Effect Diversity Is 1.2, (b) School Effect Diversity Is 0.6

Fig. 2 (a) shows that the ability rankings of examinees from a population with high school diversity have a more minor deviation from the red line compared to Fig. 2 (b), which has a larger deviation. MIRT 2PL modelling using the MML-GHQ method better predicts the ability rankings of examinees from a population with high school diversity than from a population with low school diversity. These conditions are because more homogeneous schools have fewer minor differences in ability between examinees, so errors in the estimator can distort the ability rankings of examinees.

Fig. 2 (a) shows that higher ability heterogeneity ($\tau = 1.2$) yields a larger RMSE than lower heterogeneity ($\tau = 0.6$). In contrast, Fig. 2 indicates that higher heterogeneity leads to more accurate rank-order predictions (Spearman's $\rho = 0.94$ vs. 0.82). Specifically, RMSE reflects the magnitude of estimation error on the original scale between estimated and true values. As ability heterogeneity increases (higher τ), the distribution of true abilities becomes more dispersed, with more extreme values. Under limited measurement information (e.g., nine items), posterior estimates tend to exhibit shrinkage toward the mean, which can increase deviations for extreme true values and therefore increase RMSE. In contrast, Spearman's correlation evaluates rank-order consistency and is primarily sensitive to whether relative ordering is preserved rather than to the exact numerical closeness of estimates to true values. When τ is larger, the separation in true ability across individuals (and, consequently, across schools) becomes more pronounced, making the ordering more robust to measurement noise. Thus, even if absolute scale error increases, fewer rank reversals occur, resulting in a higher Spearman's correlation.

3.3 Empirical Study

The MIRT 2PL model parameter estimates characteristics of item difficulty, item discriminability, and the mathematical literacy ability of PISA participants. The item difficulty parameter describes the value at which a PISA participant has a 50% chance of answering a question correctly. The item discriminability parameter indicates how well an item can differentiate PISA participants based on their ability levels. Table 1 shows that the MIRT 2PL model analysis results show that question number 7 is the most challenging question, followed by questions 8 and 9. These three questions have excellent item discrimination parameters.

The answers to 9 mathematics domain questions from 1302 PISA 2018 participants yielded 182 response patterns. These item response data will demonstrate Indonesian PISA participants' mathematical literacy achievement. Data analysis using the MIRT 2PL model with the MML-GHQ method yielded 1171 levels of students' mathematical literacy abilities. Fig. 3 shows the distribution of PISA participants' mathematical literacy abilities based on the MIRT 2PL model using the MML-GHQ method.

Table 1. Parameter of Item with MIRT 2PL Using The MML-GHQ Method

Item	Difficulty parameter	Discriminability parameter
1	0.00	1.00
2	1.22	1.16
3	1.10	1.84
4	1.88	0.92
5	1.83	2.46
6	2.17	0.37
7	3.35	2.12
8	2.13	2.65
9	2.13	2.04

The parameter estimation results of the MIRT 2PL model indicate that the average ability at the school level is 0.76, with a deviation between schools of 0.74 and between students within schools of 0.57. Comparing the variance between schools and the total variance in ability yields an Intraclass Correlation Coefficient (ICC) of 0.63. It means that 63% of the total variation in student ability stems from differences between schools, while approximately 37% stems from differences between students within the same school. In other words, school effects are strong in explaining differences in mathematical literacy abilities among PISA participants.

An ICC of 0.63 is exceptionally high for educational achievement data, indicating that most of the variation in mathematical literacy lies between schools rather than within schools. Substantively, this suggests strong between-school stratification, consistent with socioeconomic sorting across schools, where advantaged and disadvantaged students are concentrated in different schools. Such sorting may widen achievement gaps because schools can differ systematically in instructional quality, teacher quality and availability, facilities, learning environments, and academic support.

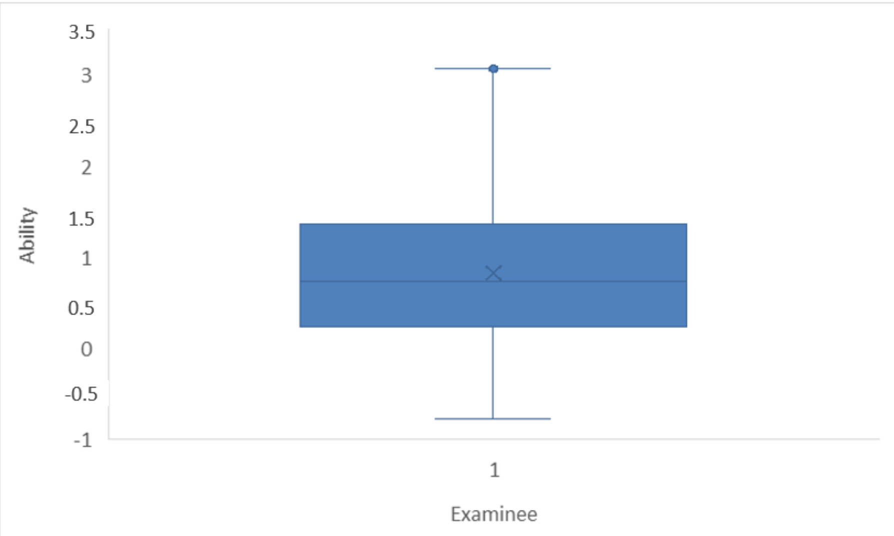


Figure 3. The Distribution of PISA Participants' Mathematical Literacy Abilities

The results of this study prove that school effects influence the mathematical literacy abilities of PISA participants. The 397 sample schools of PISA represent the mathematical literacy abilities of students from 85368 schools in Indonesia [18]. It is necessary for us to prioritize paying attention to socio-economic inequality at the school level because one study stated that socio-economic inequality is negatively correlated with student learning outcomes [21]. Equitable distribution of school resources is a significant challenge for Indonesia, so that we can improve the mathematical literacy achievements.

4. CONCLUSION

The development of the MIRT 2PL model using the proposed method (MML-GHQ) has performed well in predicting the rank order of test takers' abilities. Based on simulation results, the greater the effect of school diversity, the more accurate the prediction of the rank order of test takers' abilities. The MML-GHQ method is also efficient in terms of computation time. Increasing the number of schools, students, and test items in the empirical data produces findings that align with the simulation data. The study's results using the MIRT 2PL model with the MML-GHQ method yielded 1171 levels of students' mathematical literacy abilities. 63% of the variation in PISA participants' mathematical literacy abilities is due to school effects and 37% to student effects. Future studies will extend the MIRT 2PL model with MML-GHQ by incorporating explanatory covariates at both the student and school levels, thereby providing a more realistic representation of educational measurement conditions.

Author Contributions

Alona Dwinata: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Writing - Original Draft. Anang Kurnia: Conceptualization, Data Curation, Formal Analysis, Supervision, Software, Validation. Aji Hamim Wigena: Conceptualization, Formal Analysis, Supervision, Validation. Muhammad Nur Aidi: Conceptualization, Formal Analysis, Supervision, Validation. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

The authors thank the Organisation for Economic Co-operation and Development (OECD) as the source of the PISA data 2018.

Declarations

The authors declare that has no conflicts of interest to report study.

Declaration of Generative AI and AI-assisted Technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement, including grammar, spelling, and clarity. The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. All final text was reviewed and approved by the authors.

REFERENCES

[1] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, "GENERALIZED MULTILEVEL STRUCTURAL EQUATION MODELING," in *Psychometrika*, 2004, vol. 69, no. 2. doi: https://doi.org/10.1007/BF02295939.

[2] S. P. Reise, "ITEM RESPONSE THEORY," *Encycl. Clin. Psychol.*, pp. 1–10, 2015. doi: https://doi.org/10.1002/9781118625392.wbecp357.

[3] F. B. Baker and S.-H. Kim, "ITEM RESPONSE THEORY - PARAMETER ESTIMATION TECHNIQUES - SECOND EDITION, REVISED AND EXPANDED," *Meas. Theory Action Case Stud. Exerc.*, 2004.

[4] P. De Boeck and M. Wilson, *EXPLANATORY ITEM RESPONSE MODELS: A GENERALIZED LINEAR AND NONLINEAR APPROACH*. New York: Springer, 2004. doi: https://doi.org/10.1007/978-1-4757-3990-9

[5] R. J. Adams, M. Wilson, and M. Wu, "MULTILEVEL ITEM RESPONSE MODELS: AN APPROACH TO ERRORS IN VARIABLES REGRESSION," *J. Educ. Behav. Stat.*, vol. 22, no. 1, 1997. doi: https://doi.org/10.3102/10769986022001047.

[6] A. Kamata, "ITEM ANALYSIS BY THE HIERARCHICAL GENERALIZED LINEAR MODEL," *J. Educ. Meas.*, vol. 38, no. 1, 2001. doi: https://doi.org/10.1111/j.1745-3984.2001.tb01117.x.

[7] K. S. Maier, "A RASCH HIERARCHICAL MEASUREMENT MODEL," *J. Educ. Behav. Stat.*, vol. 26, no. 3, 2001. doi: https://doi.org/10.3102/10769986026003307.

[8] S. W. Raudenbush and A. S. Bryk, *HIERARCHICAL LINEAR MODELS: APPLICATIONS AND DATA ANALYSIS METHODS. 2ND EDITION*, vol. 1. 2002.

[9] Y. Miyazaki, Y. Chungbaek, K. O. Shropshire, and D. Hedeker, "CONSEQUENCES OF IGNORING NESTED DATA STRUCTURE ON ITEM PARAMETERS IN RASCH/1P-IRT MODEL," *Behaviormetrika*, vol. 46, no. 2, 2019. doi: https://doi.org/10.1007/s41237-019-00090-8.

[10] H. Goldstein, "MULTILEVEL MODELLING OF EDUCATIONAL DATA," in *Methodology and Epistemology of Multilevel Analysis*, 2003. doi: https://doi.org/10.1007/978-1-4020-4675-9_2

[11] J.-P. Fox, *BAYESIAN ITEM RESPONSE MODELING*. 2010. doi: https://doi.org/10.1007/978-1-4419-0742-4

[12] I. Sulis and M. D. Toland, "INTRODUCTION TO MULTILEVEL ITEM RESPONSE THEORY ANALYSIS: DESCRIPTIVE AND EXPLANATORY MODELS," *J. Early Adolesc.*, vol. 37, no. 1, 2016. doi: https://doi.org/10.1177/0272431616642328.

[13] H. Ravand, "ITEM RESPONSE THEORY USING HIERARCHICAL GENERALIZED LINEAR MODELS," *Pract. Assessment, Res. Eval.*, vol. 20, no. 7, 2015.

[14] C. König, C. Spoden, and A. Frey, "AN OPTIMIZED BAYESIAN HIERARCHICAL TWO-PARAMETER LOGISTIC MODEL FOR SMALL-SAMPLE ITEM CALIBRATION," *Appl. Psychol. Meas.*, vol. 44, no. 4, pp. 311–326, 2020. doi: https://doi.org/10.1177/0146621619893786.

[15] C. Patarapichayatham and A. Kamata, "EFFECTS OF DIFFERENTIAL ITEM DISCRIMINATIONS BETWEEN INDIVIDUAL-LEVEL AND CLUSTER-LEVEL UNDER THE MULTILEVEL ITEM RESPONSE THEORY MODEL," *Open J. Appl. Sci.*, vol. 04, no. 08, 2014. doi: https://doi.org/10.4236/ojapps.2014.48039.

[16] M. Jeon and N. Rockwood, "PLMIXED: AN R PACKAGE FOR GENERALIZED LINEAR MIXED MODELS WITH FACTOR STRUCTURES," *Applied Psychological Measurement*, vol. 42, no. 5. SAGE Publications Inc., pp. 401–402, Jul. 01, 2018. doi: https://doi.org/10.1177/0146621617748326.

[17] M. S. Johnson, "MARGINAL MAXIMUM LIKELIHOOD ESTIMATION OF ITEM RESPONSE MODELS IN R," *J. Stat. Softw.*, vol. 20, no. 10, 2007. doi: https://doi.org/10.18637/jss.v020.i10.

[18] B. Kemendikbud, "PENDIDIKAN DI INDONESIA BELAJAR DARI HASIL PISA 2018 (EDUCATION IN INDONESIA IS LEARNING FROM THE RESULTS OF PISA 2018)," *Pus. Penilai. Pendidik. Balitbang KEMENDIKBUD*, no. 021, pp. 1–206, 2019.

[19] R. D. Bock and M. Aitkin, "MARGINAL MAXIMUM LIKELIHOOD ESTIMATION OF ITEM PARAMETERS: APPLICATION OF AN EM ALGORITHM," *Psychometrika*, vol. 46, no. 4, 1981. doi: https://doi.org/10.1007/BF02293801.

[20] R. D. Bock and R. D. Gibbons, *ITEM RESPONSE THEORY*, First edit. USA: John Wiley & Sons, Inc, 2021. doi: https://doi.org/10.1002/9781119716723

[21] L. Sempé, "SCHOOL-LEVEL INEQUALITY MEASUREMENT BASED ON CATEGORICAL DATA: A NOVEL APPROACH APPLIED TO PISA," *Large-Scale Assessments Educ.*, vol. 9, no. 1, 2021. doi: https://doi.org/10.1186/s40536-021-00103-7.