Indonesian Journal of Educational Research and Technology

Journal homepage: http://ejournal.upi.edu/index.php/IJERT/

# Horizontal Equating of Science Test Forms Using Generalized Partial Credit Model (GPCM) in Secondary Education

*Fajar Nur Cahyani*, Samsul Hadi, Haryanto, Heri Retnawati*

Universitas Negeri Yogyakarta, Indonesia

*Correspondence: E-mail: fajarnur.2024@student.uny.ac.id

## ABSTRACT

Ensuring fair comparisons between different test forms is a central concern in educational assessments. This study explores the horizontal equating of two versions of a science academic test using the Generalized Partial Credit Model, which is suitable for items scored across multiple categories. Student responses were analyzed using the Mean–Sigma method, involving shared anchor items to align the scales of the two test forms. The analysis revealed consistent item parameters and student ability estimates after transformation. A good model fit was observed based on residual and approximation measures, although further refinement may be needed due to limited index values in structural comparisons. The threshold distributions became more stable after equating, and graphical analyses confirmed that item characteristics and measurement information were preserved. This approach proved successful because it aligned the measurement scales without distorting original ability estimates. The findings support the development of fairer assessment systems that uphold validity, reliability, and comparability in science education.



## ARTICLE INFO

***Article History:***
*Submitted/Received 07 Apr 2025*
*First Revised 24 May 2025*
*Accepted 17 Jul 2025*
*First Available online 18 Jul 2025*
*Publication Date 01 Dec 2025*

***Keyword:***
*Equating,*
*Generalized Partial Credit Model (GPCM),*
*Item Response Theory (IRT),*
*Mean–Sigma,*
*Polytomous Items.*

## 1. INTRODUCTION

Tests serve as a central tool in evaluating student learning outcomes, particularly within the framework of national education policies (Al Husaeni *et al*., 2024; Fiandini *et al*., 2024). In Indonesia, the use of standardized assessments follows the Regulation of the Minister of Education, Culture, Research, and Technology Number 21 of 2022, which outlines the standards for educational assessment across primary and secondary levels (see https://peraturan.bpk.go.id/Home/Details/235007/permen-kemdikbudristek-no-21-tahun-2022). To maintain test security and validity, it is common practice to construct multiple forms of a test. However, when students take different versions, the resulting scores may lack direct comparability. This inconsistency challenges the fairness and accuracy of interpretations drawn from assessment outcomes. As such, score equating becomes essential to ensure that student performance is evaluated on a common scale (Petersen, 1989; Haertel, 1986; Dorans *et al*., 2010).

Equating is a psychometric method designed to convert raw scores from one test form into the equivalent scores of another, enabling equitable interpretation. Horizontal equating, in particular, is used when two test forms are constructed from the same blueprint and competencies but administered to different groups. The importance of equating lies in its capacity to uphold justice in educational evaluations, especially when multiple test versions are used. Modern approaches to equating increasingly rely on Item Response Theory (IRT), which offers advantages over Classical Test Theory by accounting for individual item characteristics and test-taker abilities. Among various IRT models, the Generalized Partial Credit Model (GPCM) is particularly well-suited for polytomous items (those scored on multiple levels) commonly found in science assessments that involve rubrics for high-level thinking.

Despite the increasing use of IRT-based equating, most studies focus on dichotomous models such as Rasch or two-parameter logistic (2PL), with limited attention to GPCM for polytomous items (Fitriana & Soepriyanto, 2022). Research applying GPCM with horizontal equating, especially in Indonesian science education, remains scarce. Moreover, although anchor items (identical questions shared across forms) are widely recognized as essential for equating, they are often underutilized in instrument design. Because this gap limits the fairness and comparability of assessments, there is a clear need to investigate the application of GPCM in horizontally equating test forms through anchor items. This study addresses that gap by applying the GPCM and the Mean–Sigma equating method to two science test forms at the secondary level. The novelty of this study lies in integrating anchor-based horizontal equating with polytomous scoring, offering a robust method for constructing fair assessments. The findings are expected to contribute to more valid, reliable, and equitable educational measurement practices in science education.

## 2. METHODS

This study used a quantitative approach aimed at equalizing two forms of academic ability tests in the field of Natural Sciences through the application of the GPCM model. The data analysis technique used was equating analysis. Equalization was carried out horizontally, which was between two different test forms but arranged based on basic competencies and equivalent grids, and given to different groups of students. The GPCM model was chosen because it was suitable for polytomous response data, which were question items that had more than two categories of scores, such as those commonly used in assessments that measure high-level thinking skills and in-depth conceptual understanding. The general equation of GPCM is shown in equation (1).

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

$$P\left(X_{ni} = k \mid \theta_n\right) = \frac{exp\left(\sum_{m=1}^{k} a_i(\theta_n - \delta_{im})\right)}{\sum_{j=0}^{mi} exp\left(\sum_{m=1}^{j} a_i(\theta_n - \delta_{im})\right)} \quad (1)$$

where θ is the trait level, αi is the slope (i), and δij is the Intersection between lines between categories (m) per item (i).

This study adopted a quantitative research approach aimed at equating two forms of academic ability tests in the Natural Sciences domain through the implementation of the Generalized Partial Credit Model (GPCM). The primary analytical technique employed was horizontal equating, which involves comparing different test forms constructed based on identical content grids and core competencies but administered to separate groups of students. This approach ensures that each test form assesses the same constructs despite being taken by different examinees.

The GPCM was selected due to its suitability for handling polytomous response data, where items are scored across multiple ordered categories. Such scoring formats are frequently used in assessments targeting higher-order thinking and complex conceptual understanding. By accommodating these characteristics, the GPCM allows for a more accurate estimation of student ability and item parameters. The general mathematical formulation of the GPCM is presented in equation (1).

The data for this study were obtained from secondary sources, specifically the results of instrument development used in prior research. The test design followed the Non-Equivalent Anchor Test (NEAT) framework, consisting of two test forms (Form A and Form B) each comprising thirty items. Of these, eight items functioned as anchor items, identical in both content and structure across the two forms, and served as the basis for the equating procedure. Participant response data were recorded in the form of item-level scores, ranging from zero to three, following the instrument's rubric-based scoring criteria.

The population comprised Grade X students from high schools and Islamic senior secondary schools (MA) in North Sumatra Province enrolled in Natural Sciences subjects. A sample of two hundred eighty-one students was selected using a cluster random sampling technique, considering curriculum alignment and school readiness.

Data analysis proceeded in several stages. First, item parameter estimation was conducted separately for each test form using the GPCM, implemented through R software. The eight designated anchor items were then used for scale alignment through the Mean–Sigma method, a linear transformation technique that adjusts item and ability parameters based on the mean and standard deviation of the theta estimates from the anchor items. The resulting transformation coefficients (slope and intercept) were applied to the parameters and theta values of Form B, converting them to the scale of Form A.

Following equating, the adjusted item parameters and ability scores (theta) for Form B were evaluated for accuracy and consistency. Fit diagnostics and model adequacy were assessed using statistical indices such as log-likelihood (LogLik), Akaike Information Criterion (AIC), and Root Mean Square Error of Approximation (RMSEA). Additionally, the distribution of theta values before and after equating was examined to assess the effectiveness of the scale alignment. All analyses, including parameter estimation, equating, and visualization, were performed using R software. Let me know if you'd like me to refine the participants, instrument design, or analysis tools further or format this for a journal submission.

## 3. RESULTS AND DISCUSSION

In this study, two forms of academic ability test instruments in the field of science were used, namely Form A and Form B, each of which consisted of 30 questions in the form of a polytomous scale. The instruments are compiled based on a grid that has been validated by

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

experts and developed from the basic competencies of science subjects at the high school level. Both forms were given to class X students in various high schools/MA in North Sumatra Province. A total of 281 respondents participated in this trial, consisting of 141 students who worked on Form A and 140 students who worked on Form B. Respondent selection was carried out through a cluster random sampling technique, taking into account the diversity of schools based on location and accreditation status.

Of the total items, as many as 8 question items have the same structure and substance between Form A and Form B and are used as anchor items for the horizontal equalization process. Equalization is carried out by making Form A a reference scale, while Form B is equated to the scale through a mean-sigma equating approach based on the *Generalized Partial Credit Model (GPCM)* Item Response Model. The score analyzed was a polytomous score (graded score) with a value range of 0 to 3 per item. The results obtained for model fit statistics on form A and form B are shown in **Table 1**.

**Table 1**. Fit statistics of Form A and Form B models.

| Fit Statistics | Form A | Form B |
|---|---|---|
| RMSEA | 0.0286 | 0.0526 |
| SRMR | 0.081 | 0.0878 |
| TLI | 0.511 | 0.5418 |
| CFI | 0.549 | 0.578 |

Based on the results of the fit indices in **Table 1**, for the Generalized Partial Credit Model (GPCM), the RMSEA value was 0.0286 for Form A and 0.0526 for Form B. This value is below the general tolerance threshold of 0.06, which indicates a good level of model fit between the model structure and the observed data (Hu & Bentler, 1999). In absolute terms, the GPCM model is quite capable of explaining the structure of the observational data without large deviations. However, the SRMR (Standardized Root Mean Square Residual) value of 0.0878 for Form B slightly exceeds the general threshold of 0.08, which can be interpreted as a small discrepancy between the covariance matrix predicted by the model and the observed covariance matrix (Bentler, 1990). Meanwhile, the values of TLI (Tucker-Lewis Index) and CFI (Comparative Fit Index) only reached 0.511 and 0.549 for Form A, and 0.5418 and 0.578 for Form B, respectively. These values are well below the recommended reference value of ≥ 0.90 (Hooper *et al.*, 2008), which suggests that the GPCM model in this context is still not optimal in capturing the diversity of data structures relative to null models.

Thus, although RMSEA showed a good match and SRMR was still within the tolerable range, the low TLI and CFI values suggest that the model used may not yet fully represent the complexity of participant response data. Therefore, improvements to the model or re-checking the data structure and assumptions of the GPCM model need to be carried out to improve the quality of parameter estimation and the accuracy of participant score interpretation. Furthermore, the results of the estimation of Form A and Form B parameters were obtained, which were analyzed using *the Generalized Partial Credit Model (GPCM)* with the mean-sigma equating method. The results are shown in **Table 2**.

The results obtained in **Table 2** showed that of the 30 Form A items, most had low to moderate discrimination values. However, there are a number of items with negative discrimination values and extreme thresholds, such as A9, A13, and A28, indicating that these items do not function well psychometrically. It is necessary to validate the content of these problematic items to consider revision or item elimination. Meanwhile, the estimated parameters for form B are shown in the following **Table 3**.

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

**Table 2**. Estimation of Form A parameters using GPCM.

| Items | a | b1 | b2 | b3 |
|---|---|---|---|---|
| A1 | 0.445 | 3.157 | 0.735 | -1.001 |
| A2 | 0.484 | 1.992 | -1.356 | -0.206 |
| A3 | 0.404 | -2.737 | -0.843 | -5.518 |
| A4 | 0.222 | 0.158 | 6.407 | 2.670 |
| A5 | 0.222 | 8.335 | -5.422 | 4.561 |
| A6 | 0.112 | 15.022 | -8.322 | -4.688 |
| A7 | 0.318 | 0.463 | -0.561 | -1.069 |
| A8 | 0.118 | 11.285 | 0.456 | 17.818 |
| A9 | -0.021 | -8.952 | -75.053 | -25.707 |
| A10 | 0.197 | 3.462 | 0.803 | -1.320 |
| A11 | 0.347 | 0.144 | 3.937 | -1.986 |
| A12 | 0.137 | 0.311 | 21.561 | -15.833 |
| A13 | -0.068 | -19.742 | 8.325 | -21.002 |
| A14 | 0.393 | 0.909 | -2.728 | -1.844 |
| A15 | 0.123 | -0.587 | 5.682 | 0.159 |
| A16 | 0.177 | 9.765 | -6.833 | 13.035 |
| A17 | -0.091 | -25.622 | 11.231 | -17.102 |
| A18 | 0.355 | 1.172 | 2.822 | -1.514 |
| A19 | 0.068 | 18.886 | 4.071 | 6.846 |
| A20 | -0.232 | -8.421 | 3.821 | -2.324 |
| A21 | 0.340 | 1.801 | 2.571 | 1.485 |
| A22 | 0.026 | 44.654 | 23.946 | 52.806 |
| A23 | 0.300 | 9.245 | -1.519 | -3.427 |
| A24 | 0.272 | 3.160 | 0.620 | -5.373 |
| A25 | 0.253 | -1.413 | 4.093 | -2.258 |
| A26 | 0.170 | 2.958 | 7.890 | -1.026 |
| A27 | 0.518 | 0.541 | -1.854 | -0.452 |
| A28 | 0.010 | 146.100 | 4.460 | 199.454 |
| A29 | -0.120 | -3.908 | -5.303 | -5.629 |
| A30 | -0.121 | -18.291 | 5.241 | -20.395 |

Based on the results shown in **Table 3**, the results of the parameter analysis of Form B items show that most items have a fairly adequate level of discrimination, but there are some items with negative discrimination and extreme thresholds that indicate potential anomalies in the response data. This can be caused by an imbalance in the category of scores or the quality of the questions that need to be improved. Overall, the results obtained in **Tables 2 and 3** are summarized in **Table 4**.

**Table 4** shows the descriptive results of the threshold parameter (b) of the items in Form A and Form B, which have been equalized through *the Mean–Sigma Equating* method in the *Generalized Partial Credit Model (GPCM) model,* it is found that there is a difference in the distribution of difficulty levels between the two forms of questions. The average threshold in Form A is 4.021 with a standard deviation of 24.041, while in Form B (after equating), the average threshold drops to 0.509 with a standard deviation of 8.283. This decrease shows that the equating process successfully calibrates the difficulty level of Form B items to be on the same scale as Form A so that it is more proportionate and representative of the test-takers' general ability.

The threshold value (b) should ideally be spread around zero to reflect items of moderate difficulty and spread evenly across the participant's ability spectrum. Thresholds that are too extreme (e.g., exceeding ±10) can indicate a disruption in the estimation process, such as

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

unused response categories or unvaried answer patterns. In this case, the threshold range on Form A reaches a minimum value of -36.571 to a maximum of 116.671, which is far beyond the reasonable limit and indicates the possibility of an extreme item or an estimated anomaly. After the equating process, the threshold range on Form B became more manageable (-25,760 to 28,133), although some values still showed extreme estimates that needed to be explored further.

**Table 3**. Estimation of Form B parameters using GPCM.

| Items | A | b1 | b2 | b3 |
|---|---|---|---|---|
| B23 | 0.192 | -4.601 | -0.125 | -5.614 |
| B24 | 0.376 | 3.462 | -0.118 | -0.240 |
| B25 | 0.387 | 4.746 | -3.474 | 2.716 |
| B26 | -0.154 | -13.168 | -0.184 | |
| B27 | 0.190 | 7.774 | -10.433 | 11.886 |
| B28 | 0.381 | -3.562 | 6.557 | -7.262 |
| B29 | -0.186 | -11.521 | 4.917 | -18.528 |
| B30 | 0.054 | 14.996 | 10.455 | 16.409 |
| B31 | 0.149 | 13.962 | -18.489 | 12.708 |
| B32 | 0.407 | 4.116 | -4.161 | 4.773 |
| B33 | 0.438 | -1.474 | 4.026 | -4.706 |
| B34 | 0.475 | 1.006 | 1.377 | -1.275 |
| B35 | 0.573 | -0.265 | 0.728 | -0.781 |
| B36 | 0.283 | -4.312 | 11.033 | -5.417 |
| B37 | 0.299 | 3.873 | 2.844 | -1.749 |
| B38 | 0.603 | -1.297 | 0.071 | -1.382 |
| B39 | 0.489 | -1.653 | 3.116 | -4.664 |
| B40 | 0.576 | -1.868 | 2.816 | -5.452 |
| B41 | 0.316 | 3.414 | 2.725 | 1.378 |
| B42 | -0.089 | -33.314 | -4.744 | 9.080 |
| B43 | 0.343 | 3.507 | -2.673 | 2.518 |
| B44 | 0.485 | 2.996 | -0.462 | 2.683 |
| B45 | 0.061 | 11.292 | 21.251 | -7.672 |
| B46 | 0.245 | -1.024 | 5.336 | -6.056 |
| B47 | 0.369 | 2.299 | -1.886 | 4.386 |
| B48 | 0.439 | 0.967 | 0.632 | 1.327 |
| B49 | -0.045 | -39.374 | 6.093 | -44.000 |
| B50 | 0.022 | 69.352 | -4.832 | 19.878 |
| B51 | 0.069 | 17.047 | -15.430 | 6.905 |
| B52 | 0.233 | 3.985 | 6.492 | -4.189 |

**Table 4**. Summary of Form A and Form B that Equating has done.

| Form | Mean_b | SD_b | Min_b | Max_b |
|---|---|---|---|---|
| A | 4.021 | 24.041 | -36.571 | 116.671 |
| B (Equated) | 0.509 | 8.283 | -25.760 | 28.133 |

Large standard deviations in the threshold distribution indicate inconsistencies in the level of difficulty between items, which in the context of test development can have an impact on the reliability and fairness of the assessment. Therefore, although the equating results have significantly improved the threshold distribution, further evaluation of items showing negative discrimination values or extreme thresholds is still necessary to ensure the validity

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

of the measurement instrument. The following results of the threshold (b) on form B before and after equating are presented in **Table 5**.

**Table 5**. Comparison of the Threshold before and after equating.

| Items | b before equating | B Equating |
|---|---|---|
| B23 | -2.828 | -2.869 |
| B24 | -0.177 | -0.179 |
| B25 | -0.374 | -0.379 |
| B26 | -0.091 | -0.092 |
| B27 | 0.715 | 0.727 |
| B28 | -0.348 | -0.352 |
| B29 | -6.707 | -6.806 |
| B30 | 13.236 | 13.432 |
| B31 | -2.849 | -2.890 |
| B32 | 0.301 | 0.306 |
| B33 | -0.335 | -0.340 |
| B34 | 0.050 | 0.051 |
| B35 | -0.027 | -0.027 |
| B36 | 2.767 | 2.808 |
| B37 | 0.539 | 0.548 |
| B38 | -0.647 | -0.656 |
| B39 | -0.763 | -0.774 |
| B40 | -1.299 | -1.318 |
| B41 | 2.021 | 2.051 |
| B42 | 2.136 | 2.168 |
| B43 | -0.076 | -0.077 |
| B44 | 1.094 | 1.110 |
| B45 | 6.690 | 6.789 |
| B46 | -0.355 | -0.360 |
| B47 | 1.231 | 1.250 |
| B48 | 0.965 | 0.979 |
| B49 | -18.677 | -18.953 |
| B50 | 7.413 | 7.523 |
| B51 | -4.201 | -4.262 |
| B52 | 1.134 | 1.152 |

Based on the comparison of the threshold value (*b*) in Form B before and after the *equating* process using the Mean–Sigma method in the Generalized Partial Credit Model (GPCM), it can be concluded that there has been a subtle but consistent scale adjustment. Almost all items experienced b-value changes in a very small range, averaging only about ±0.01 to ±0.2. For example, item B25 goes from –0.374 to –0.379, and item B31 from –2.849 to –2.890. These changes show that the equating process does not drastically change the difficulty characteristics of the items, but rather manages to align the positions of the items so that they are on the same measurement scale as Form A.

Adjustments like this reflect the good *equating* principle, which is that changes in scores should be minimal but systematic to maintain the psychometric integrity of the test. The main goal of equating is so that two different forms of testing can be used equally without creating bias against a particular group of respondents. In this context, the equating method successfully ensures that the b-value of Form B still reflects the item's original difficulty level, but has been adjusted to the reference scale of Form A. Thus, the equating process carried out not only produces stable and accurate technical results but also guarantees the accuracy of measurements across test forms. This is very important in the context of competency-

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

based educational assessments that prioritize equal opportunities for every student in facing various forms of exam questions. The following also presented a visualization of the ability (theta) of students who worked on form B before and after equating. The visualization is shown in **Figure 1**.

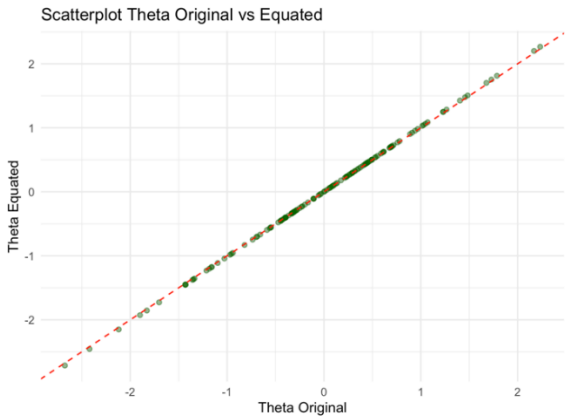



**Figure 1.** Ability (θ) of students before and after equating.

**Figure 1** shows the distribution of the relationship between theta scores before and after equalization using the Mean–Sigma method. Most of the points are around the identity line, indicating that the equating process results in a commensurate scale of ability between the two forms. This shows that the equalization process is effective and does not damage the structure of the participants' initial ability.

Furthermore, the Test Information Function (TIF) will be displayed for each form of test, Form A and Form B. The comparison of the results of TIF Form A and Form B can be seen in **Figure 2**. Based on **Figure 2**, the TIF curve of Form B has a higher peak of information compared to Form A, especially in the intermediate ability range (θ≈0). This shows that Form B is more efficient in measuring the ability of participants who are around the average population. The higher information peaks reflect that the estimated capabilities of Form B will have a smaller standard error rate than Form A in that range.

In addition, the curve pattern of both test forms shows a symmetrical distribution, indicating that the item design is evenly distributed across the capability spectrum, but Form B contributes more information to the distribution center area. The higher the value of information at a point θ, the more accurate the estimation of ability at that point. Thus, these results confirm that Form B has higher measurement reliability for participants with average ability, while Form A provides more limited information.

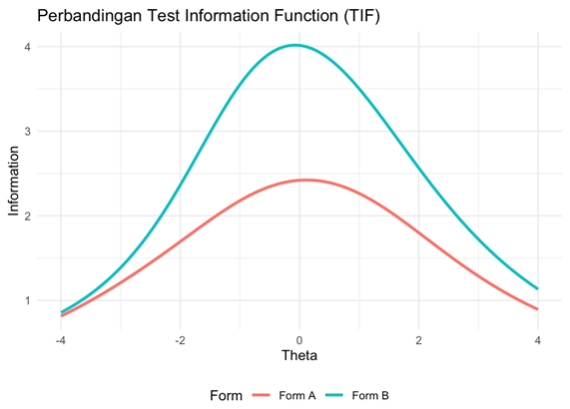
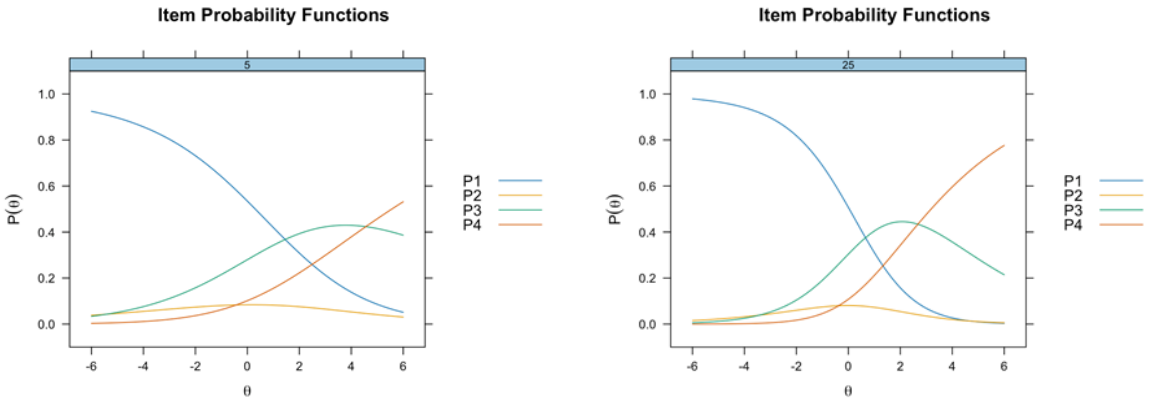



**Figure 2.** The comparison of the results of TIF Form A and Form B.

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

**Figure 3a** shows the Item Characteristic Curve (ICC) for Item 5 on Form A, while Figure 4 shows the ICC for Item 25, which serves as the anchor item. Each curve on the graph represents the probability of selection of a particular score category (P1, P2, P3, P4) by participants with different ability levels (θ). The P1 curve (lowest category) shows a high probability value at low ability (θ <−2), and decreases as the ability increases, as expected. In contrast, the P4 curve (the highest category) increases with θ, reaching a peak in participants with high ability. The P2 and P3 curves show a fairly smooth and intersecting probability of transitions between categories, indicating that these items have an orderly and functional categorization structure.

In **Figure 3b**, the ICC anchor pattern of item 25 also shows good characteristics with the probability distribution between the categories spread proportionally along the range of capabilities. It was seen that P2 and P3 had peaks around θ = 0 to 2, while P4 increased sharply in participants with high ability (θ > 2). This pattern indicates that the anchor item works effectively to distinguish participants with different skill levels. Overall, **Figures 3a and 3b** show the ICC of two question items with a categorization level of 4. The transition between categories is gradual and responsive to changes in participants' abilities (θ). This indicates that both Item 5 (Form A) and Item 25 (Anchor Form B) have optimal and discriminatory measurement functions, as recommended in the GPCM model principle.




(a) (b)

**Figure 3.** (a) ICC Form A Item 5; (b) ICC Anchor Item 25.

## 4. CONCLUSION

This study shows that the application of the GPCM with a horizontal equating approach using the Mean–Sigma method can be used effectively to equate two forms of academic tests in Natural Sciences subjects. The equalization process through eight anchor items resulted in a valid and stable scale transformation, demonstrated by the consistency of theta scores before and after equalization as well as the distribution of the items' parameters that were logical and did not undergo significant distortion.

The value of the equating constant obtained is slope = 0.9794 and intercept = 0.0294. After the equating process, the distribution of the threshold value (b) of Form B was successfully calibrated to be on the same scale as Form A (mean b: Form A = 4.021; Form B equated = 0.509). The information function curve (TIF) shows that Form B has a higher information power than Form A, especially in the medium capability range, while the ICC shows that most items have a corresponding transition pattern between categories. Despite some items with negative discrimination values or extreme thresholds need to be reviewed to ensure that the

DOI: https://doi.org/10.17509/ijert.v5i3.88050
p- ISSN 2775-8419 e- ISSN 2775-8427

measurement quality remains optimal. The model statistical fit (RMSEA < 0.06) indicates a good model fit, although the low CFI and TLI values indicate that the model can still be improved. Overall, this approach has been shown to be able to maintain the validity and reliability of the instrument between two different forms of testing, and can be recommended for the development of polytomous score-based assessments in the context of formal education.

## 5. ACKNOWLEDGMENT

We would like to express our sincere gratitude to the supervisors and academic advisors who provided valuable guidance throughout the completion of this research. Special thanks are extended to the schools and students in North Sumatra who participated in the data collection process. We also acknowledge the prior work of Tri Yanti Nadapdap, whose instrument development served as the foundation for this study.

## 6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. Authors confirmed that the paper was free of plagiarism.

## 7. REFERENCES

Al Husaeni, D.F., Al Husaeni, D.N., Fiandini, M., and Nandiyanto, A.B.D. (2024). The research trend of statistical significance test: Bibliometric analysis. *ASEAN Journal of Educational Research and Technology, 3*(1), 71-80.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246.

Dorans, N. J., Moses, T. P., and Eignor, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series*, *2010*(2), i-41.

Fiandini, M., Nandiyanto, A.B.D., Al Husaeni, D.F., Al Husaeni, D.N., and Mushiban, M. (2024). How to calculate statistics for significant difference test using SPSS: Understanding students comprehension on the concept of steam engines as power plant. *Indonesian Journal of Science and Technology*, *9*(1), 45-108.

Fitriana, Y., and Soepriyanto, Y. (2022). Implementasi model IRT 2PL dalam penyetaraan nilai ujian sekolah. *Jurnal Penelitian dan Evaluasi Pendidikan, 26*(1), 13–25.

Haertel, E. (1986). The valid use of student performance measures for teacher evaluation. *Educational Evaluation and Policy Analysis*, *8*(1), 45-60.

Hooper, D., Coughlan, J., and Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53–60.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Petersen, N. S. (1989). Uses and misuses of standardized tests. *The Phi Delta Kappan, 70*(8), 634–639.