JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v15i1.40842
http://journal.uinjkt.ac.id/index.php/jp3i

Investigating Differential Item Functioning (DIF) in Geometry Test Scores: Holistic vs. Analytical Scoring Rubrics

Raoda Ismail[1], Okky Riswandha Imawan[1], Heri Retnawati[2], Haryanto[2]

Mathematics Education, Universitas Cenderawasih, Papua, Indonesia[1]
Educational Research and Evaluation, Universitas Negeri Yogyakarta, DIY, Indonesia[2]

raodaismail26@gmail.com

Abstract

The use of polytomous data in test instruments enables more detailed assessment of test-takers' abilities, but group differences, such as gender, class, and ethnicity, are often overlooked. Differential Item Functioning (DIF) analysis helps determine whether these identities influence test performance. This descriptive quantitative study examines DIF in geometry tests using Holistic and Analytical Scoring Rubrics across gender, class, and ethnic groups. The study involved 102 undergraduate students from Cenderawasih University, Papua, who completed a geometry test with 10 descriptive questions. Two scoring rubrics were used: the Holistic Scoring Rubric with three categories and the Analytical Scoring Rubric with five. Results were analyzed with the difR package in the R Program. The criterion for detecting DIF is a p-value less than 0.05. Findings show that some test items exhibit DIF concerning gender, class, and ethnicity. The DIF detected for the holistic scoring rubric is items 1, 6, 7, and 10 for the gender group; items 1, 4, 5, and 8 for the class group; and items 1, 2, and 6 for the ethnic group. Meanwhile, the DIF detected for the analytic scoring rubric is item 10 for the gender group; items 9 and 10 for the class group; and item 10 for the ethnic group. However, not all DIF items are flawed; some assess fundamental skills. The Analytical Rubric demonstrated slightly higher reliability (alpha 0.903) than the Holistic Rubric (alpha 0.804). These insights support the development of more equitable and sustainable assessment practices, ensuring fairness and inclusivity in educational evaluations.

Keywords: Analytical Scoring, DIF, Geometry Assessment, Holistic Scoring, Polytomous Data

Abstrak

Penggunaan data politomus dalam instrumen tes memungkinkan penilaian kemampuan peserta tes yang lebih rinci, tetapi perbedaan kelompok seperti gender, kelas, dan etnisitas sering diabaikan. Analisis Differential Item Functioning (DIF) membantu menentukan apakah identitas-identitas ini mempengaruhi kinerja tes. Studi kuantitatif deskriptif ini meneliti DIF dalam tes geometri menggunakan Rubrik Penilaian Holistik dan Analitik pada kelompok gender, kelas, dan etnis. Studi ini melibatkan 102 mahasiswa sarjana dari Universitas Cenderawasih, Papua, yang menyelesaikan tes geometri dengan 10 soal uraian. Dua rubrik penilaian digunakan: Rubrik Penilaian Holistik dengan tiga kategori dan Rubrik Penilaian Analitik dengan lima kategori. Hasil dianalisis dengan paket difR dalam Program R. Kriteria butir soal yang mengandung DIF yaitu ketika hasil analisisnya menunjukkan p-value kurang dari 0,05. Temuan menunjukkan bahwa beberapa item tes menunjukkan DIF terkait gender, kelas, dan etnis. DIF terdeteksi untuk rubrik penskoran holistik yaitu butir 1, 6, 7, dan 10 di kelompok gender; butir 1, 4, 5, dan 8 di kelompok kelas, serta butir 1, 2, dan 6 di kelompok etnis. DIF terdeteksi untuk rubrik penskoran analitik yaitu butir 10 di kelompok gender; butir 9, dan 10 di kelompok kelas, serta butir 10 di kelompok etnis. Namun, tidak semua item DIF dianggap cacat; beberapa item menilai keterampilan dasar. Rubrik Analitik menunjukkan reliabilitas yang sedikit lebih tinggi (alpha 0,903) dibandingkan dengan Rubrik Holistik (alpha 0,804). Wawasan ini mendukung pengembangan praktik penilaian yang lebih adil dan berkelanjutan, memastikan keadilan dan inklusivitas dalam evaluasi pendidikan.

***Kata kunci**: Data Politomus, DIF, Penilaian Analitik, Penilaian Geometri, Penilaian Holistik*

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Introduction

This study investigates Differential Item Functioning (DIF) in a geometry test using polytomous data, examining whether gender, class, and ethnicity influence test outcomes. Conducted with 102 undergraduate students from Cenderawasih University, the research employs both Holistic and Analytical Scoring Rubrics to evaluate responses to 10 descriptive geometry items. By applying descriptive quantitative methods, the study seeks to identify potential biases in test items and explore how different scoring rubrics affect the detection of DIF. The goal is to enhance the fairness and inclusivity of educational assessments, particularly in evaluating the mathematical competencies of prospective primary school teachers.

Geometry tests are essential tools for measuring spatial reasoning and problem-solving skills among students in teacher education. These assessments use both theoretical and practical problems, scored using either holistic or analytical rubrics. Holistic rubrics assess overall performance, while analytical rubrics evaluate specific response components, potentially influencing DIF detection. The distinction between these rubrics is especially relevant when considering how polytomous data—Graded under holistic scoring and Partial under analytical scoring—can affect the identification of item bias. Recognizing how scoring methods interact with demographic variables is vital to ensuring that assessments accurately reflect students' abilities, free from unfair advantages or disadvantages.

Tests with polytomous items often provide more detailed information about a test taker's proficiency (Nandakumar, R., Feng Yu, Hsin-Hung Li, 1998). Research indicates varying opinions on the optimal number of scoring categories for maximizing internal consistency reliability. Some researchers advocate for 7 scales, while others suggest 4 or even 3 scales (Ismail et al., 2024; Lei Chang, 1994; Santoso et al., 2023). Linn and Gronlund recommend using a 3-7 score scale for effective measurement (Boughton, K.A., Klinger, D.A., & Gierl, 2001). Several studies highlight that polytomous response analysis can improve measurement accuracy, suggesting that scoring development should increasingly utilize the polytomous scoring system with multiple categories (Baker, J. G., Rounds, J. B., & Zeron, 2000) and (Imawan et al., 2025; Tognolini, J., & Davidson, 2003).

A major concern in geometry testing is accurately and consistently scoring assessments, especially when considering potential bias introduced by different scoring rubrics. Holistic rubrics assess students' responses based on an overall impression, which can introduce subjectivity and allow extraneous factors—like demographic characteristics—to influence scores. This subjectivity may lead to inconsistent evaluations across student groups and contribute to Differential Item Functioning (DIF). Conversely, analytical rubrics break down responses into specific components, enabling a more structured, detailed evaluation that tends to increase scoring consistency. However, even analytical rubrics can unintentionally favor certain groups depending on how their criteria are defined or weighted, making both scoring methods relevant in the analysis of DIF.

This study examines how DIF appears in geometry test results scored with holistic versus analytical rubrics, particularly among prospective primary school teachers. By comparing these scoring methods, the research aims to identify how each may contribute to demographic bias, including gender and ethnic differences. Holistic rubrics, though quicker to apply, offer only general performance feedback, while analytical rubrics provide detailed insight into individual strengths and weaknesses. Understanding how each rubric influences DIF detection is essential for promoting fairness in educational assessments. The findings are expected to inform better practices in educational measurement and enhance the validity and equity of assessments used in teacher training programs.

In addition to the instrument's validity, reliability, and item characteristics, another crucial aspect to consider in assessment is DIF. DIF analysis can identify items in the instrument that are advantageous to specific groups of test participants, such as those based on gender. Research has shown the benefits of using DIF analysis, such as a study of Indonesian national examination results, which identified 36 items exhibiting DIF, indicating that these items were beneficial only to certain groups of test takers (Setiawan et al., 2024).

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

In conducting a measurement, a valid and reliable measuring instrument is essential. Such an instrument ensures that re-measurement results accurately reflect what is intended to be measured, free from extraneous factors. If a test is affected by factors other than those being measured, it is said to exhibit test bias (Aziz & Günther, 2023; Retnawati, 2014). Furthermore, test items should not be discriminatory. If a test contains items that favor a particular group, it is said to exhibit DIF, a bias. DIF item bias is defined as the difference in the probability of answering correctly between two groups, often referred to as the Focal group and the Reference group (Angoff, 1993). According to Zumbo (1999), DIF occurs when test-takers from two groups with the same ability have different probabilities of answering correctly, indicating a bias. In unidimensional item response theory, DIF is expressed as the difference in the probability of a correct response between the Focal and Reference groups. Test bias arises when decisions based on test scores unfairly impact one group more than another (Osterlind & Everson, 2009).

There are two interpretations of item bias: Social and Statistical. Social bias pertains to the assessment and use of items, while statistical bias refers to DIF. When researchers use the term "bias," it typically refers to the social meaning, whereas "DIF" refers to statistical bias (Angoff, 1993). An item may exhibit DIF but still be considered unbiased if it measures a specific ability aligned with the measurement objectives. In such cases, one group might have a weakness in that specific area, even if they possess the same overall ability. According to Hambleton et al (1991), a test exhibits Differential Item Functioning (DIF) when there is a difference in scores caused by elements that are either beneficial or detrimental to certain participants. DIF occurs when different groups with the same ability obtain different expected scores on the same item. DIF is a key concept in measuring bias (Sheppard, R., 2006), involving systematic error and affecting validity. Additionally, individuals within groups may contribute to errors, potentially impacting reliability when they respond to items exhibiting DIF.

Hortensius (2012) identifies the source of DIF as related to group membership, such as differences in social and economic class, area of residence, race, gender, region, culture, and ethnicity. Sumintono and Widhiarso (2013) suggest that individual bias arises from different performances across items. For example, when researchers administer the same test in different ways (e.g., conventional vs. modern), DIF may emerge. In relation to construct validity, Messick (1995) found that irrelevant constructs are a primary source of bias in interpreting test scores. Irrelevant construct variance can cause items to function differently across groups, leading to DIF. Furthermore, Sumintono and Widhiarso (2013) argue that DIF can occur when an item favors individuals with specific characteristics, while disadvantaging those with opposing traits. For instance, an item in a children's test might involve drawing a snowflake to identify discrepancies. This task is straightforward for children familiar with snow, but challenging for those who have never encountered it.

One of the key benefits of Item Response Theory (IRT) in assessment instrument analysis is its capacity to assess the performance level or category attained by test participants based on their responses. This enables the collection of detailed information regarding item characteristics and the estimation of participants' abilities using an ability scale (Otaya et al., 2020). IRT can also be applied to evaluate the quality of test items, for example, through Rasch analysis. If certain items are found to be unsuitable after this analysis, problematic items are flagged, reviewed, and revised by experts in the field to ensure they meet the necessary standards for future assessments. This process ensures that the item bank, containing these revised items, remains both psychometrically and theoretically robust, making it an effective resource for developing assessment instruments (Yim et al., 2024).

In IRT, the number of participants in a test instrument's trial can influence the accuracy of the item characteristic analysis. Research has shown that the 2PL model was identified as the most suitable, with RMSD values indicating that the number of test participants affects the stability of item parameter estimation using the 2PL model (Ibrahim et al., 2024). However, IRT analysis requires strict conditions, particularly regarding sample size when testing an instrument. Some studies have found no significant differences between IRT and classical analysis results; for example, research has shown that item analysis using both Classical Test Theory and Item Response Theory approaches did not reveal notable differences in the difficulty index across Packages 1-5 of math tests (Kartowagiran et al., 2019).

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Test item characteristics, including the difficulty index and discriminatory power, can be analyzed through Item Response Theory (IRT). Research has shown that IRT can effectively measure both the difficulty index and discriminatory power of test items, as well as the content area being assessed (Karimah et al., 2021). Some studies have explored more complex parameter models to deepen understanding of IRT, particularly emphasizing the importance and interpretation of the guess parameter, an additional component of the model. The Carelessness parameter reflects the probability that a test participant will answer an item correctly based on their genuine knowledge, suggesting that the participant truly knows the correct answer. This parameter accounts for atypical responses from high-ability test takers, which may result in the guess parameter being less than 1. In the Item Characteristic Curve (ICC), the Carelessness parameter is represented by the height of the asymptote at the top of the curve, which indicates the maximum likelihood that a high-ability participant will correctly answer a specific item (Pardede et al., 2023).

An advantage of using IRT analysis is that research findings show it can accurately categorize test items into difficulty levels—such as difficult, medium, or easy—and assess each item's discriminatory power. The results of this analysis can also be used to evaluate the effectiveness of questions and provide insight into how well the test accommodates participants with varying levels of ability (Kusumawati & Hadi, 2018). Additionally, IRT analysis can identify which items fall into the good or poor categories. It can also reveal the types of mistakes participants make while answering questions, allowing educators to focus on strengthening the content areas where students face challenges (Kartianom & Mardapi, 2018).

The analytical models for polytomous data include the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM), which are defined by two main parameters: Difficulty Level (b) and Discriminative Power (a). The primary difference between these models is that, in GPCM, the difficulty index of each category need not follow a strict order. This means that the difficulty of category k might exceed that of category k + 1. For example, an item in a descriptive test with an analytical scoring rubric might have a first step that is more challenging than the second. Similarly, the probability of answering correctly in category k could be higher than in category k - 1. Tang (1996) explains that in GPCM, the probability that examinees select category k is defined as the difference between the probabilities of those with scores greater than or equal to k and those with scores greater than or equal to k + 1. In contrast, GRM assumes that the difficulty index for each category is ordered, so higher categories correspond to higher difficulty levels. Therefore, if a respondent answers a step correctly, they are expected to have answered all preceding steps correctly as well.

Research indicates that both GRM and GPCM are effective for scoring polytomous tests. Hambleton and Swaminathan (2013) found that GPCM more accurately reflects how data is generated, resulting in a Test Information Curve (TIC) that provides more precise ability estimates than GRM. Furthermore, studies on the form of description instruments and testlets reveal that: (1) Empirically and through simulations, tests in the form of descriptions have a higher item information function than testlets; (2) Simulations indicate that item number and sample size affect the comparison of item information function values between descriptions and testlets; and (3) GRM modeling is most accurate with 20 items and a sample size of 2000 (Susongko, 2010).

In polytomous scoring, instrument makers have the freedom to use various categories. Some research findings include: (1) Partial Credit Model scoring with weighting provides more accurate ability estimates than non-weighted or dichotomous scoring; (2) Increasing the number of categories in partial credit scoring enhances accuracy (Wasis, 2011). Additionally, student learning outcomes vary depending on the scoring guidelines used. Research findings indicate: (1) In mathematics, there are 47 underlying attributes, including 4 content, 36 process, and 7 skill attributes; (2) The highest incompleteness of content, process, and skill attributes is in geometry and measurement; (3) The highest error type in numbers, algebra, geometry, and measurement is conceptual, while in statistics and probability, it is language interpretation; and (4) Diagnostic information on Mathematics UN data can be obtained through attribute identification, developing a polytomous scoring rubric, calculating attribute incompleteness, and diagnosing test taker errors (Isgiyanto, 2013).

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Based on theoretical and empirical evidence, both Holistic and Analytical scoring methods in polytomous assessments have unique strengths and weaknesses that significantly affect the quality of test outcomes. The holistic rubric, using a simpler three-category scale (0, 1, 2), offers a broad evaluation of student responses but may be subject to bias and reduced sensitivity to subtle performance differences, potentially allowing demographic biases to go undetected. In contrast, the analytical rubric, with its five-category scale (0–4), evaluates specific aspects such as conceptual understanding, reasoning, and accuracy. This detailed approach provides more precise feedback and a greater potential to detect Differential Item Functioning (DIF), as it is more sensitive to variations in test-taker performance across demographic groups. Therefore, the choice of scoring method is critical to ensuring fair and reliable assessments, and research comparing these two approaches is essential to optimize scoring accuracy and minimize bias.

This study aims to analyze how scoring rubrics influence the detection of DIF in geometry assessments for prospective primary school teachers. It includes comparing the reliability and fairness of three-category holistic and five-category analytical scoring models. The research steps involve conducting DIF analysis, assessing the reliability of each rubric, and making recommendations for improved scoring practices. While holistic rubrics may obscure demographic disparities due to their generality, analytical rubrics offer a more nuanced view of performance and are better equipped to detect item bias. Ultimately, this research will contribute valuable insights into how rubric design affects assessment equity, helping to develop more inclusive and valid evaluation systems within educational measurement.

Methods

This study employs a descriptive quantitative research design. The objectives are: (1) to describe the results of the DIF analysis for the Holistic rubric geometry test across gender, class, and ethnic groups, and (2) to describe the results of the DIF analysis for the Analytical rubric geometry test across the same groups. The results of this study are explained using both classical and modern test theories. Classical test theory was chosen, given the limited number of study participants, with only 102 students. Although the sample size is relatively small, classical test theory remains relevant because it provides a straightforward, easily understandable overview of test results, making it suitable for data that do not require complex statistical techniques. However, the primary reason for using classical theory is to provide a clear basis for comparison with newer test theories.

On the other hand, modern test theory was employed to support and further develop the results obtained from classical theory. The use of modern test theory, such as Differential Item Functioning (DIF) analysis, offers advantages for detecting test bias related to factors such as gender, class, and ethnicity. This theory enables more detailed and in-depth analysis, which cannot be achieved by classical theory alone. Through DIF analysis, this study identifies test items that may function differently for different groups, thereby enhancing the overall accuracy and validity of the test.

The combined use of both theories provides a more comprehensive and balanced approach to analyzing test results. Classical test theory offers a solid foundation for general understanding of test performance, while modern test theory enriches the analysis by offering deeper insights into fairness and bias in evaluation, ultimately leading to more accurate and inclusive assessments.

To ensure that the instrument is unidimensional and well-constructed, six key components of IRT-based analysis were evaluated: unidimensionality, local independence, item parameter estimation, reliability (person and item separation), Wright Map, and the Test Information Function. The analysis was conducted using Item Response Theory models, specifically the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM).

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Participants

The subjects of this study were 102 undergraduate students from the Elementary School Teacher Education Study Program at Cenderawasih University, Papua, Indonesia. Although Scott et al. (2009) recommended a minimum of 300 participants per group for robust DIF analysis, the sample size of 102 participants in this study remains valuable. Several studies have shown that, in exploratory or pilot studies, smaller sample sizes can still make meaningful contributions to the development of theory and practice, as long as the research design is carefully constructed and the results are interpreted with caution (Andrade, 2020; Ploutz-Snyder et al., 2014). Research involving a limited number of participants is also common in education, especially when researchers work with specific, restricted populations, such as students from a particular study program in a specific region.

Andrade (2020) emphasizes that in exploratory social research, large sample sizes are not always required to obtain valuable initial insights. Even in quantitative studies, small samples can still be effectively analyzed with appropriate analytical approaches, such as repeated-measures designs or the use of valid and reliable instruments. Furthermore, Ploutz-Snyder et al. (2014) explain that small-scale studies can be highly valuable, particularly when the research population is limited or difficult to access, and when the study is designed to generate a preliminary understanding of a phenomenon. McGrath and Brandon (2018) from Advances in Neonatal Care emphasized that small-scale studies are often sufficient as pilot studies, intended to test feasibility, instruments, and procedures, without the intention of making broad generalizations.

Instruments

Data collection was conducted using a descriptive geometry test comprising 10 questions, which had been content-validated by experts in elementary school mathematics education.

Data Collection

While larger sample sizes are typically preferred to improve statistical power, this study serves as a preliminary or exploratory analysis into DIF detection in geometry assessments. Despite the reduced statistical power, the DIF detection methods employed using the difR package remain effective for identifying DIF in this sample. Therefore, the study's findings should be seen as contributing valuable insights into potential biases in geometry assessments, especially for underrepresented groups. This research provides practical value for educators and test developers working in similar contexts, and while the results are limited in generalizability, they offer a solid foundation for future investigations with larger sample sizes.

In this study, the participants were divided into three distinct groups based on key demographic variables: gender, classroom (A, B, C, D, E), and ethnicity. The purpose of this grouping was to investigate potential Differential Item Functioning (DIF) across these subgroups, helping to identify whether test items functioned differently for distinct groups with equivalent ability levels. The gender grouping consisted of male and female participants. This classification aimed to explore whether there were any gender-related biases in the test items, which might cause differential performance despite similar underlying abilities. Gender differences in educational assessments can sometimes result from cultural, social, or psychological factors, making it essential to ensure that test items are fair and unbiased for all genders.

The second grouping was based on classroom, with participants categorized into Classroom A, B, C, D, and E. This classification was important for examining potential biases related to classroom conditions, teaching methods, or other environmental factors that could affect students' test performance. By comparing the two classrooms, this study sought to determine whether the same test items displayed varying levels of difficulty or discrimination for students in different classrooms, even when their abilities were presumed to be similar. The third grouping was based on ethnicity, distinguishing between Papuan and non-Papuan students. This distinction was crucial for assessing potential ethnic biases in the test items. Papua, a culturally diverse region, may have students with diverse educational experiences and

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

backgrounds, which could influence how they respond to certain test items (Handoko et al., 2023; Ismail et al., 2024). Understanding ethnic-based DIF is important for ensuring that the assessment tool is valid and fair across diverse cultural groups, so that all students, regardless of their ethnic background, are evaluated on equal terms.

By dividing the participants into three groups—gender, classroom, and ethnicity—the study aimed to investigate how test items functioned differently across these subgroups. This approach helps to ensure that the geometry test, used for prospective elementary school teachers, accurately and equitably measures the intended competencies without bias towards any particular group. The findings of the study provide valuable insights into the fairness of the test and can guide improvements in test design to ensure that all students are assessed in a manner that is both valid and equitable.

Data Analysis

The geometry test results, evaluated using both the Holistic and Analytical scoring rubrics, were analyzed using the difR package in R. In this study, the difR package was used to detect Differential Item Functioning (DIF) in the geometry test data. The difR package provides several statistical methods, including Mantel-Haenszel, logistic regression, and IRT-based techniques, to identify items exhibiting DIF across groups (e.g., gender, ethnicity). DIF occurs when individuals from different groups with the same underlying ability have different probabilities of answering a particular item correctly, indicating potential test bias. To detect DIF, the difR package computes a p-value for each test item. The criterion for identifying DIF is a p-value less than 0.05 for an item. A p-value below this threshold suggests that the item behaves differently across groups, indicating the presence of DIF. Items with p-values greater than or equal to 0.05 are considered non-biased and exhibit no significant DIF. The DIF analysis was conducted using a dataset that included responses to the geometry test along with a grouping variable (e.g., gender or ethnicity). The results of the DIF analysis helped identify which items might favor one group over another, ensuring the fairness and validity of the assessment.

To determine the item parameters, the Graded Response Model (GRM) and Generalized Partial Credit Model (GPCM) analyses were employed, focusing on the Difficulty Level (b) and Discriminatory Power (a) parameters. In this study, GRM and GPCM analyses were conducted to determine the item parameters of the geometry test. These models, implemented using the ltm and mirt packages, are based on Item Response Theory (IRT) and are suitable for analyzing polytomous items, where responses can vary in degree rather than a simple correct/incorrect format. The primary focus was on two key parameters: Difficulty Level (b) and Discriminatory Power (a). The Difficulty Level (b) indicates how challenging an item is, with higher values suggesting more difficulty. The Discriminatory Power (a) measures how well an item differentiates between individuals with different ability levels. Items with higher discriminatory values are more effective at distinguishing between test-takers of varying abilities. These IRT analyses provide insights into the performance of each test item, ensuring that the items are both valid and reliable in measuring the intended constructs. By evaluating the difficulty and discriminatory parameters, this study aimed to refine the assessment and enhance the accuracy of the geometry test.

Results and Discussion

Results

The results of this study are described classically and modernly. The classical theory was used because the number of subjects in this study was not too large, namely, only 102 students. Meanwhile, modern theory is used in this research to support the results of classical theory and to develop it, as it can produce more detailed results using the DIF analysis model.

Unidimensionality

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

The assumption of unidimensionality is a fundamental prerequisite in Item Response Theory (IRT), ensuring that item responses are primarily driven by a single latent trait or construct (Embretson & Reise, 2000). In the present study, unidimensionality was examined using the Confirmatory DETECT procedure developed by Stout (1990) and implemented in the R package sirt (Robitzsch, 2025). This approach provides a nonparametric method for assessing dimensionality directly from item response data, without relying on assumptions of exploratory factor analysis. Unlike traditional factor analysis, which is based on linear relationships among observed variables, the DETECT framework assesses local item dependencies and conditional covariances within the IRT framework, thus aligning with the requirements of polytomous IRT models such as the GRM and the GPCM.

In this analysis, the total sum score of respondents was used as the conditioning variable, and items were grouped into clusters based on their item labels (A1–A10). The DETECT procedure computes several indices, including DETECT, ASSI (Approximate Simple Structure Index), RATIO, MADCOV, and MCOV, which together provide evidence for or against the presence of multidimensionality. Specifically, DETECT measures the degree of conditional covariance between item clusters: values near zero (or below 0.20) indicate essential unidimensionality, whereas larger positive values (e.g., > 1.0) suggest the existence of multiple latent dimensions (Strout, 1990; Zhang, 2006). The ASSI index reflects the degree to which items align with a single dominant dimension, whereas the RATIO index represents the proportion of explained covariance accounted for by the detected structure.

The results obtained in this study were as follows: DETECT = −5.838, ASSI = −0.422, and RATIO = −0.551 (both weighted and unweighted values were identical). According to the interpretive guidelines proposed by Stout (1990) and further elaborated by Zhang (2006), DETECT values substantially below 0.20 and negative ASSI and RATIO values indicate that the observed covariances among item pairs are minimal and largely random, implying that a single latent trait adequately explains response patterns across items. This outcome strongly supports the assumption of unidimensionality.

These findings imply that the set of items measuring mathematical literacy operates along a single latent dimension—consistent with the theoretical construct underlying the assessment. Consequently, the data meet the unidimensionality assumption required for the application of IRT models such as the GRM and the GPCM. This confirms that each item's parameters (difficulty and discrimination) can be validly interpreted in relation to a common underlying ability continuum. Moreover, because the DETECT approach is robust to moderate local dependencies and non-normal response distributions, the result enhances the validity of using IRT-based modeling in this study.

Local Independence

Local independence was evaluated using Yen's Q3 statistic. The residual correlations between item pairs ranged from -0.45 to 0.19. Since none of the Q3 values exceeded the commonly accepted threshold of 0.20, there was no evidence of local item dependence. This finding supports the assumption that item responses are conditionally independent, given the latent trait.

Item Measure/Item Parameters (Difficulty Level (b) and Discriminatory Power (a))

The GRM and GPCM analysis models can be considered equivalent because both are able to analyze up to two parameters, namely the level of difficulty (b) and discriminating power (a) for polytomy data. The following are the results of item parameters using GRM and GPCM for the use of Holistic and Analytical Scoring Rubrics.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

GRM and GPCM for Holistic Scoring Rubric (3 categories)

Table 1. Parameters of GRM and GPCM Items for Holistic Scoring Rubric

Item	GRM			GPCM		
	a	b1	b2	a	b1	b2
H1	1.55244	-1.76613	0.59677	14.09987	-17.3253	0.55072
H2	1.91010	-1.22557	-0.31189	14.89492	-0.92412	-0.65905
H3	0.78990	-4.37070	-0.63751	0.683732	-38.2124	-0.84855
H4	0.78098	-2.08173	-0.72173	0.551837	-0.44193	-21.8933
H5	1.33855	-1.50045	0.64222	10.94660	-14.3125	0.54578
H6	2.27062	-0.92287	-0.17901	16.85529	-0.63541	-0.47486
H7	0.89267	-1.87462	0.78532	0.676253	-16.3183	0.62460
H8	1.60900	-1.79720	-0.44493	12.92257	-15.3504	-0.68671
H9	2.55409	-1.76327	-0.26504	23.16839	-17.4282	-0.30482
H10	1.68423	-2.09471	-0.49931	14.62149	-19.2316	-0.64617

Sources: Personal Data (2025).

Table 1 GRM shows that the discriminatory parameter (a) spreads from 0.789 to 2.554, which means that the items on the geometry test are able to distinguish the test takers who are less capable and those who are smart. Meanwhile, most of the geometry test items are in the Very Easy to Moderate category, seen from the Difficulty Level (b), which ranges from -4.37 (b1 question number 3) to 0.785 (b2 question number 7). It can be explained from Table 4 GPCM that the Discriminatory Power (a) spreads from 0.552 to 2.317, which means that the items on the geometry test are able to distinguish the less capable test takers from the smart ones. Meanwhile, most of the geometry test items are in the Very Easy to Moderate category, seen from the Difficulty Level (b), which ranges from -3.821 (b1 question number 3) to 0.625 (b2 question number 7).

GRM for Analytical Scoring Rubric (5 categories)

Table 2. GRM Item Parameters for Analytical Scoring Rubric

Item	a	b1	b2	b3	b4
A1	1.999931	-1.5735527	-1.2978564	-0.2875215	0.7786220
A2	2.238653	-1.2732061	-1.0765742	-0.5590903	0.1102500
A3	2.136511	-2.1931462	-1.4016972	-0.8618440	0.6159733
A4	1.681197	-1.3831069	-1.1235018	-0.7577617	0.6405128
A5	2.077074	-1.5920856	-0.5716486	-0.2247910	0.8687028
A6	3.448846	-0.9629615	-0.6484354	-0.4944125	0.4736772
A7	1.564446	-1.7692878	-0.9665331	0.1568280	1.0182997
A8	2.806392	-1.6066616	-1.1512352	-0.7068113	0.6191394
A9	3.513294	-1.7188991	-0.8118338	-0.4470420	0.6787220
A10	1.447032	-3.2493874	-1.9683735	-1.4088194	-0.5296398

Sources: Personal Data (2025).

Table 2 shows that the Discriminatory Power (a) spreads from 1.447 to 3.513, which means that the items on the geometry test are able to distinguish the test takers who are less capable and those who are smart. Meanwhile, most of the geometry test items are in the Very Easy to Medium category, as indicated by the Difficulty Level (b), which ranges from -3.249 (b1, question number 10) to 1.018 (b4, question number 7).

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

GPCM for Analytical Scoring Rubric (5 categories)

Table 3. GPCM Item Parameters for Analytical Scoring Rubric

Item	a	b1	b2	b3	b4
A1	1.2898338	-0.28115907	-1.8963756	-0.3962280	0.5175347
A2	1.3237489	0.08664854	-1.5864877	-0.7581657	-0.3604084
A3	1.4417502	-1.76320672	-0.9942148	-1.2350889	0.5911886
A4	0.9352642	0.44630137	-1.2219840	-1.7958353	0.5663788
A5	1.1375030	-1.55229343	0.2176705	-0.9587152	0.7433489
A6	2.3212297	-0.59231429	-0.3107159	-1.0761423	0.4072492
A7	0.7579220	-1.09921642	-1.3234467	0.6999405	0.3444077
A8	1.7910214	-1.16491201	-0.9883288	-1.029.921	0.6346376
A9	2.3517799	-1.70166852	-0.5480829	-0.7245770	0.6537364
A10	0.8381363	-2.72120115	-0.8239045	-1.5963524	-1.3947628

Sources: Personal Data (2025).

Table 3 shows that the Discriminatory Power (a) spreads from 0.758 to 2.351, which means that the items on the geometry test are able to distinguish the test takers who are less capable and those who are smart. Meanwhile, most of the geometry test items are in the Very Easy to Moderate category, seen from the Difficulty Level (b), which ranges from -2.721 (b1 item number 10) to 0.743 (b4 item number 5).

Reliability (Person & Item Separation Reliability)

The Graded Response Model converged after 68 iterations, yielding a final log-likelihood of -1180.526 and a maximum parameter change of 0.00010, indicating stable convergence. The marginal reliability of the instrument was 0.91, suggesting a high degree of precision in estimating individuals' latent trait levels.

Table 4. Estimation of Classical Reliability: Holistic and Analytical Scoring Rubric

Cronbach's alpha for the data-set	
Item	10
Sample units	102
Alpha Holistic Scoring (3 categories)	0.804
Alpha Analytical Scoring (5 categories)	0.903

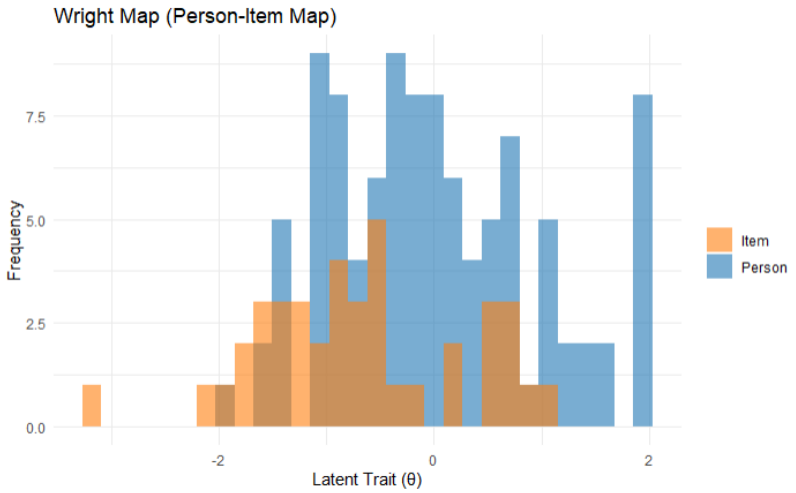
Sources: Personal Data (2025).

The results of data processing show that the Cronbach's alpha coefficient for the Holistic Scoring Rubric is 0.804, which is high. The results of data processing show that the Cronbach's alpha coefficient for the Analytical Scoring Rubric is 0.903, which is very high. The following is the DIF analysis using Holistic and Analytical scoring rubrics, combined with the grouping of Gender, Class, and Ethnicity among geometry test participants.

Wright Map

The Wright Map in Figure 2 demonstrates good alignment between person abilities and item thresholds. Item difficulties span the range of respondent abilities with minimal gaps, indicating effective targeting of the instrument. The overlapping distributions confirm that the test provides a reliable measurement across the latent trait continuum.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

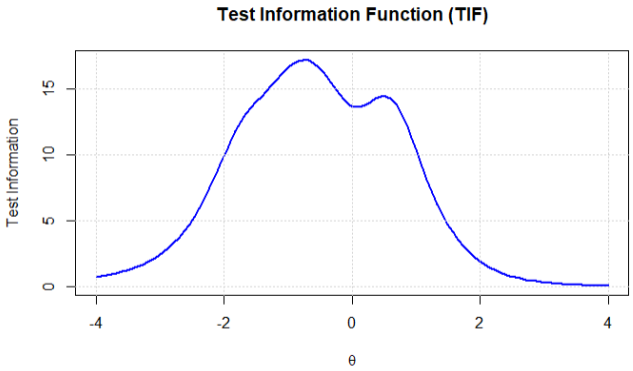


Sources: Personal data (2025).

Figure 1. Wright Map

Test Information Function

The Test Information Function (TIF) in Figure 3 demonstrates that the instrument provides high measurement precision across a wide range of the latent trait. Peak information occurs in the central portion of the trait continuum, indicating that the test is most informative for respondents with average to slightly above-average ability levels. Overall, the TIF confirms that the test reliably differentiates individuals across relevant levels of the trait.



Sources: Personal data (2025).

Figure 2. Test Information Function

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

DIF Analysis for Holistic Rubrics in Gender Grouping, Class Grouping, and Ethnic Grouping

Table 5. DIF Output for Holistic Rubric in Each Grouping

Item	*Gender Grouping*			*Class Grouping*			*Ethnic Grouping*		
	Stat.	**P-Value**	**DIF**	**Stat.**	**P-Value**	**DIF**	**Stat.**	**P-Value**	**DIF**
H1	7.96	0.004	Yes	5.768	0.016	Yes	7.164	0.007	Yes
H2	0.252	0.618	-	1.260	0.261	-	5.435	0.019	Yes
H3	0.253	0.618	-	1.162	0.264	-	1.894	0.168	-
H4	2.352	0.126	-	9.160	0.002	Yes	1.883	0.154	-
H5	1.304	0.255	-	31.247	0.000	Yes	0.403	0.525	-
H6	5.653	0.018	Yes	0.271	0.602	-	9.702	0.001	Yes
H7	5.622	0.018	Yes	1.557	0.212	-	0.062	0.803	-
H8	0.364	0.549	-	9.652	0.001	Yes	0.026	0.870	-
H9	0.131	0.723	-	0.220	0.638	-	0.000	0.987	-
H10	5.953	0.015	Yes	0.009	0.921	-	0.197	0.657	-

Sources: Personal Data (2025).

In Table 5, the items detected as DIF are the use of Holistic rubrics for items 1, 6, 7, and 10. This means that participants of different genders, even though their abilities are equal, will have different probabilities of answering items 1, 6, 7, and 10 correctly. In Table 2, the items detected as DIF are the use of Holistic rubrics for items 1, 4, 5, and 8. This meant that participants from different classes, even though their abilities were equal, had different probabilities of answering items 1, 4, 5, and 8 correctly. In Table 2, the items detected as DIF are the use of Holistic rubrics for items 1, 2, and 6. This means that participants from different ethnic groups, even though their abilities are equal, will have different probabilities of answering items 1, 2, and 6 correctly.

DIF Analysis for Analytical Rubrics in Gender Grouping, Class Grouping, and Ethnic Grouping

Table 6. DIF Output for Analytical Rubric in Each Grouping

Item	*Gender Grouping*			*Class Grouping*			*Ethnic Grouping*		
	Stat.	**P-Value**	**DIF**	**Stat.**	**P-Value**	**DIF**	**Stat.**	**P-Value**	**DIF**
H1	0,142	0705	-	0.031	0.859	-	0.142	0.705	-
H2	0,234	0,622	-	0.125	0.723	-	1.285	0.256	-
H3	0,200	0,654	-	0.123	0.721	-	0.200	0.654	-
H4	0,777	0,377	-	0.020	0.887	-	0.665	0.412	-
H5	1,185	0,276	-	0.153	0.695	-	0.667	0.414	-
H6	0,333	0,563	-	0.051	0.819	-	0.000	1.000	-
H7	0,055	0,813	-	0.603	0.437	-	0.056	0.811	-
H8	0,123	0,765	-	0.290	0.589	-	2.000	0.157	-
H9	0,333	0,563	-	4.888	0.027	Yes	0.200	0.654	-
H10	9,000	0,000	Yes	9.123	0.000	Yes	8.000	0.004	Yes

Sources: Personal Data (2025).

In Table 6, the items detected as DIF are the use of the Analytical rubric for item 10. This means that participants of different genders, even though their abilities are equal, will have different probabilities of answering item 10 correctly. In Table 3, the items detected as DIF are the use of the Analytical rubric for items 9 and 10. This means that participants from different classes, even though their abilities are equal, will have different probabilities of answering items 9 and 10 correctly. In Table 3, the items detected as DIF are the use of the Analytical rubric for item 10. This means that participants from different ethnic groups, even though their abilities are equal, will have different probabilities of answering item 10 correctly.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Discussion

The following discussion provides a more in-depth analysis of the research results, along with supporting and refuting evidence from similar studies.

Classical Reliability Discussion

Table 7. Discussion of Classical Reliability

Use of Scoring Rubric	Cronbach's Alpha Coefficient	Category
Holistic (3 categories)	0.804	High
Analytics (5 categories)	0.903	Very High

Sources: Personal Data (2025).

Table 7 shows that the classical reliability estimate for the Analytical Scoring Rubric is higher than that for the Holistic Scoring Rubric. The Analytical Scoring Rubric, which uses five categories, has a Cronbach's Alpha coefficient of 0.903, which is categorized as "Very High." In contrast, the Holistic Scoring Rubric, with three categories, has a Cronbach's Alpha of 0.804, classified as "High." A high-quality instrument is one that has undergone testing procedures, such as the Cronbach Alpha reliability test (Wardani et al., 2018). This result indicates that the Analytical Scoring Rubric provides a more accurate assessment of respondents' abilities, reflecting greater internal consistency. The higher reliability of the Analytical Scoring Rubric corroborates the findings of Baker et al (2000), Tognolini and Davidson (2003), and Wu (2003), which suggest that multiple-category scoring is more effective than dichotomous scoring for measuring test-takers' abilities. These studies highlight the advantage of utilizing multiple categories to capture the nuances of respondents' performances more effectively.

Discussion of Modern Reliability

Table 8. Discussion of Modern Reliability

Use of Scoring Rubric	GRM	GPCM
Holistic (3 categories)	Suitable for abilities from about -3.7 to 1.8	Suitable for abilities from about -3.5 to 1.7
Analytics (5 categories)	Suitable for abilities from about -3.6 to 2	Suitable for abilities from about -2.3 to 1.5

Sources: Personal Data (2025).

The combination of the Analytical Scoring Rubric with the GPCM in Table 8 shows the narrowest range, indicating that this combination results in less error. This finding also suggests that the Analytical Scoring Rubric, paired with the GPCM, provides a more precise estimate of test-takers' abilities within the specified range, from -2.3 to 1.5. The holistic rubric for GRM, which comprises three scoring categories, spans an ability range of approximately -3.7 to 1.8. This indicates that the holistic rubric is sufficiently sensitive for assessing examinees with low to moderate abilities but may be less effective for those with higher ability levels. In contrast, the analytic rubric, which includes five scoring categories, slightly extends the measurable range toward higher abilities, from -3.6 to 2. This broader range suggests that the analytic rubric provides greater discriminatory power, particularly for examinees with moderate to high abilities, due to its more detailed categorization. Although the difference is not substantial, the analytic rubric demonstrates a slight advantage in capturing higher ability levels when used with the GRM. Nonetheless, both rubrics are effective for assessing examinees within the low-to-moderate ability range, depending on the assessment goals and focus.

Additionally, the curve depicted in Table 8 supports these findings by showing that the Analytical Scoring Rubric with GPCM yields the most information, as evidenced by the high peak of the curve exceeding theta of 25. A larger information function corresponds to a more accurate estimation, as highlighted by Hambleton et al (1991) and Lin (2008). The high peak of information in this combination indicates that the Analytical Scoring Rubric with GPCM effectively captures the nuances of test-takers'

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

abilities, minimizing measurement errors and enhancing reliability. These results are specific to the geometry test used in this study and may not necessarily apply to other tests or disciplines. As such, the findings should be interpreted within the context of this specific assessment. Different subjects or typeest tests might yield different results, underscoring the need for further research to generalize these findings across diverse contexts.

DIF Discussion

DIF is important to conduct during the selection process to determine which test items are suitable for use. As a recommendation, DIF detection should also consider other factors such as gender, school location, and other variables that could potentially cause DIF in the test items (Setiawan et al., 2024).

Table 9. Discussion of DIF

Use of Scoring Rubric	Grouping of Test Participants		
	Gender	**Class**	**Ethnic**
Holistic (3 categories)	Items indicated by DIF are numbers 1, 6, 7, and 10	Items indicated by DIF are numbers 1, 4, 5, and 8	Items indicated by DIF are numbers 1, 2, and 6
Analytics (5 categories)	The item indicated by DIF is number 10	Items indicated by DIF are numbers 9 and 10	The item indicated by DIF is number 10

Sources: Personal Data (2025).

DIF was detected in the holistic scoring rubric with the gender group, including items 1, 6, 7, and 10, as described in Table 9. Based on data analysis, the average female score on items 1, 6, 7, and 10 is higher than the average male score, making these items more favorable to females than to males. DIF was detected in the holistic scoring rubric with the class group, including items 1, 4, 5, and 8. Based on data analysis, the average score for class C on items 1 and 5 is higher than that of classes A, B, D, and E, indicating that these items are more favorable to class C than to the other classes. Based on data analysis, the average score for class A on items 4 and 8 is higher than that of classes B, C, D, and E, indicating that these items are more favorable to class A than to the other classes. DIF was detected in the holistic scoring rubric with the ethnic group in items 1, 2, and 6. Based on data analysis, the average scores for items 1, 2, and 6 for non-Papuan ethnic groups are higher than those for Papuan ethnic groups, indicating that these items are more favorable to non-Papuan ethnic groups than to Papuans.

DIF was detected in the analytical scoring rubric with the gender group in item 10. Based on data analysis, the average score for females on item 10 is higher than that for males, indicating that this item is more favorable to females than to males. DIF was detected in the analytical scoring rubric with the class group, items 9 and 10. Based on data analysis, the average score for class A on item 9 is higher than for classes B, C, D, and E, making this item more favorable to class A than to the other classes. Based on data analysis, the average score for class C on item 10 is higher than for classes A, B, D, and E, making this item more favorable to class C than to the other classes. DIF was detected in the analytical scoring rubric for item 10, with the ethnic group. Based on data analysis, the average score for non-Papuan ethnic groups on item 10 is higher than that for Papuan ethnic groups, indicating that this item is more favorable to non-Papuan ethnic groups than to Papuans.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Table 10. Item with DIF Detected in Analytical Scoring Rubric

Item	Question	DIF	Reason
9	A company is designing a cube-shaped storage container to ship fragile items. The container has edges that measure 9 cm on each side. However, the company wants to add a protective foam layer around the inside of the container to prevent the items from getting damaged. If the foam layer is 1 cm thick on each side, what will the new volume of the container be after the foam is added?	Item more favorable to class A than the other classes	This problem is more favorable to Class A because it involves advanced mathematical thinking, such as understanding how the addition of a foam layer affects the volume of a cube. Class A students, likely being more advanced, are better equipped to handle multi-step problems, apply real-world contexts to geometry, and think critically about how changes in dimensions impact 3D objects. The problem's complexity and the need for problem-solving skills align well with the curriculum and capabilities of Class A, making it more accessible and engaging for them compared to other classes.
10	In a design workshop for a school project, students are tasked with creating different shapes for a geometric garden. They need to incorporate at least 6 types of quadrilaterals and 6 types of triangles to design the layout for the garden's paths and flower beds. Your task is to draw 6 different types of quadrilaterals and 6 different types of triangles that could be used in the design of the garden.	Item more favorable to females than males. Item more favorable to non-Papuan ethnic groups than to Papuans	This item may be more favorable to females and non-Papuan ethnic groups due to cultural and educational factors. Females, often socialized to focus on design, aesthetics, and organization, might find tasks involving creativity and geometric design more engaging. Non-Papuan students, particularly those from urban areas, may have more exposure to structured education in geometry and design, making them more familiar with such tasks. In contrast, Papuans, especially those from rural areas, might have less access to such educational opportunities, and the task may not align as closely with their cultural practices or prior experiences.

Sources: Personal Data (2025).

As previously explained, Differential Item Functioning (DIF) indicates potential bias in an item. Therefore, items that show DIF should undergo a more in-depth analysis to determine whether they require revision or should be excluded from the test. However, not all items that exhibit DIF are necessarily flawed. For instance, an item may represent a fundamental skill or concept in a specific field of knowledge, such as calculating the area of a square in geometry. Even if DIF is detected, such an item might still be retained because it measures a basic skill that all individuals, regardless of gender, class, or ethnicity, must master.

In Table 9, it can also be observed that, for the Holistic Rubric, item 1 shows DIF by gender, class, and ethnicity. Despite this, the item remains relevant because it assesses a basic geometry skill that studentsall mastered should universally master. Therefore, item number 1 will not be revised. However, it is essential to investigate why gender, class, and ethnicity differences may influence the likelihood of correctly answering this item. Additionally, the sample's data distribution should be included to better understand these factors, such as the percentage of male and female participants and ethnic group representation, to provide context for the DIF findings.

One example of a solution is that mathematics lecturers must treat students according to their circumstances, because students cannot be generalized, as each may have different backgrounds, characteristics, and learning needs. Each student has a unique learning style. Some students may grasp material more easily through visual aids, such as images or diagrams, while others may understand better through verbal explanations or hands-on practice. If a lecturer uses only one teaching method without considering these differences in learning styles, some students might struggle to comprehend the material. Hence, lecturers must adapt their teaching approaches to accommodate various learning styles, ensuring that all students have an equal opportunity to understand the content.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Moreover, even though students are at the same level of education, their understanding of a concept may differ. Some students may quickly grasp the material, while others may need more time to understand it. For this reason, lecturers need to be flexible in adjusting the pace of instruction and providing additional support to students who need it, such as offering extra explanations or supplementary learning materials. This ensures that every student receives the help they need to master the material without leaving anyone behind. Gender differences can also affect how students interact during learning. For instance, female students may be more inclined to work in groups or collaborate, while male students may feel more comfortable with an individual approach. Although there is no difference in intellectual ability between men and women, their ways of interacting and contributing to class may vary. Lecturers should ensure that all students, both male and female, are treated equally and given the same opportunities to demonstrate their abilities. They should also avoid gender biases or stereotypes and create an inclusive environment for all students.

Furthermore, students' educational or social backgrounds can influence how they understand and absorb material. Students from schools with a strong foundational understanding of mathematics may be better prepared to grasp more complex concepts, while students from backgrounds with less depth in mathematics might need simpler explanations or reinforcement of basic concepts first. Lecturers must be sensitive to these differences and provide extra attention to students who may need additional help to catch up and fully comprehend the material. Ethnic diversity also plays an essential role in the learning process. Students from different ethnic backgrounds may have different approaches to learning. For example, students from one ethnic group may be more accustomed to structured, formal learning, while those from another may feel more comfortable with open-ended methods and discussions.

Lecturers should be aware of all these differences and strive to adjust their teaching methods to ensure they are acceptable to all students, regardless of their ethnic backgrounds. Overall, mathematics lecturers must be able to recognize and appreciate the diversity within their classrooms. Therefore, it is crucial for mathematics lecturers to treat each student individually and tailor teaching methods to their specific circumstances and needs. Assuming that all students learn the same way or have equivalent levels of understanding is an unfair oversimplification, as each student may have different learning styles and speeds. By understanding that each student brings unique experiences, backgrounds, and learning styles, lecturers can create a fairer and more inclusive learning environment. This allows all students, without exception, to have an equal opportunity to understand and master the material being taught.

The same thing happened to item number 10 for the use of the Holistic rubric. This item is also an item that measures the basic ability of geometry, namely "types of triangles and quadrilaterals", so it must be mastered by every student without exception, regardless of gender, class, and ethnicity. So that it is not revised item number 1, but needs to be evaluated why differences in gender, class, and ethnicity can affect the probability of answering item number 10 correctly. DIF refers to a situation in which test takers with comparable abilities, but from different demographic groups, have varying probabilities of obtaining the same outcome. A DIF study demonstrates that items free from DIF exhibit stronger construct validity (Imawan et al., 2024; Ismail et al., 2024; Sumin et al., 2022).

Discussion of Fit Items

Table 11. Discussion of Fit Items

Use of Scoring Rubric	GRM	GPCM
Holistic (3 categories)	All items fit	All items fit
Analytics (5 categories)	All items fit	All items fit

Sources: Personal Data (2025).

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

From Table 11, it is evident that all items fit perfectly with both the Graded Response Model (GRM) and the Generalized Partial Credit Model (GPCM), regardless of whether the Holistic or Analytical Scoring Rubric was used. This finding highlights the robustness of the test items in accurately capturing the underlying abilities of the test-takers. The successful fitting of all items in both models indicates that the test items are well-constructed and can effectively differentiate among varying levels of test-takers' abilities. This aligns with the statement by Dodeen (2004), which suggests that a high discriminatory parameter (a) enhances the likelihood of items fitting well during calibration with the GPCM. In this study, the geometry test items demonstrated high discriminatory power, contributing to their optimal fit.

The remarkable fit of all items in this study is particularly noteworthy given the relatively small sample size of 102 students. Modern analytical models like GRM and GPCM typically require larger sample sizes to achieve accurate item calibration. For instance, a study by Susongko (2010) found that GRM modeling was most accurate with a sample size of 2000 and 20 test items, highlighting the challenges faced when working with smaller samples. A study reports that the GPCM model is the best fit, as it has the lowest AIC and BIC values, suggesting a lower error rate (Harsana et al., 2024). Not all assessment tools use dichotomous scoring, which is why various item analysis techniques exist to support polytomous scoring. A study that focused on items with polytomous scoring sought to explore the benefits of using the MCM/GPCM model compared with the GRM model in a mixed-item format for Mathematics assessments. The findings revealed that the MCM/GPCM model yielded more precise estimations than the GRM model (Abadyo & Bastari, 2015).

This study's use of the Graded Response Model (GRM), Generalized Partial Credit Model (GPCM), and Differential Item Functioning (DIF) analysis is supported by recent empirical studies. Mutmainna et al. (2024) applied Rasch, GRM, and GPCM models to analyze the Indonesian version of the Colorado Learning Attitude Science Survey, demonstrating the effectiveness of these models for psychometric evaluation in the Indonesian educational context. Additionally, Yudiana, Triwahyuni, and Susanto (2023)employed a multidimensional Rasch model to detect gender-based DIF in the Indonesian Collective Intelligence Test–High (TIKI-T), revealing biased items in certain subtests. Findings from these studies provide empirical support that the use of GRM, GPCM, and DIF analysis is both relevant and applicable in educational research settings in Indonesia, including those with limited sample sizes but carefully designed instruments and methodologies.

Conclusion

This study's findings are specifically applicable to the geometry test questions developed by the researchers, which were validated by experts. This study's grouping of test-takers was limited to Gender, Class, and Ethnicity, suggesting a need for further DIF analysis across other groupings, such as previous school background, socioeconomic status, and more. In conclusion, the excellent fit of all items in this study highlights the effectiveness of the test design and the suitability of both GRM and GPCM for modeling the geometry test used. The findings reinforce the importance of item quality and discriminatory power in achieving reliable assessment outcomes. As research continues in this area, attention to sample size and item construction will be vital in further enhancing the accuracy and validity of educational assessments.

The analysis of Differential Item Functioning (DIF) across various demographic groups reveals significant patterns in both holistic and analytical scoring rubrics. In the holistic rubric, gender differences were evident in items 1, 6, 7, and 10, with females consistently scoring higher, indicating that these items favored females over males. Class-related DIF was observed in items 1, 4, 5, and 8, where class C performed better on items 1 and 5, while class A outperformed others on items 4 and 8. Ethnic differences were also detected in items 1, 2, and 6, with non-Papuan ethnic groups scoring higher than Papuans. In the analytical rubric, gender DIF was found in item 10, where females scored higher than males. Class-related DIF was identified in items 9 and 10, with class A performing better on item 9, while class C had a higher average score on item 10 compared to other classes. Ethnic DIF was detected in item 10, where non-Papuan ethnic groups scored higher than Papuans. These findings suggest that certain items may be

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

biased towards specific groups, highlighting the need for careful consideration in item design and scoring to ensure fairness across all demographic categories.

Although this study used a relatively small sample of 102 participants, which is below typical recommendations for GRM, GPCM, and DIF analyses, the sample is considered adequate given the exploratory nature and specific population. Small sample sizes may affect the stability of parameter estimates and the sensitivity to detect DIF. However, by using valid instruments and rigorous procedures, this study minimizes these limitations. Prior research suggests that small samples in exploratory contexts can still provide valuable preliminary insights, which should be confirmed with larger samples in future studies (Andrade, 2020; McGrath & Brandon, 2018; Ploutz-Snyder et al., 2014).

Acknowledgment

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

Conflict of Interest

The authors declare that they have no conflicts of interest.

Author Contribution

Conceptualization, design, data collection, analysis, and writing, R.I.; conceptualization, providing technical support, & supervision, R.I., O.R.I., and H.R.; conceptualization, design, providing technical support, & supervision, S.; design & providing technical support, O.R.I. and H.; design & providing technical support, O.R.I and R.I.. All authors have read and agreed to the published version of the manuscript.

References

Abadyo, & Bastari. (2015). Estimation of ability and item parameters in mathematics testing by using the combination of 3plm/grm and mcm/gpcm scoring model 1. In *Research and Evaluation in Education Journal e* (Vol. 1, Issue 1). http://journal.uny.ac.id/index.php/reid

Andrade, C. (2020). Sample Size and its Importance in Research. *Indian Journal of Psychological Medicine*, *42*(1), 102–103. https://doi.org/10.4103/IJPSYM.IJPSYM_504_19

Angoff, W. H. (1993). *Perspective on Differential Item Functioning Methodology. Holland, P.W & Wainer, H.(eds.). Differential Item Functioning.* NJ: Lawrence Erlbaum Associates, Publishers.

Aziz, R., & Günther, U. (2023). Psychometric Properties of Creative Personality Scale among Secondary School Students. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, *12*(2), 162–176. https://doi.org/10.15408/jp3i.v12i2.31808

Baker, J. G., Rounds, J. B., & Zeron, M. A. (2000). A comparison of graded response and rasch partial credit models with subjective well-being. *Journal of Educational and Behavioral Statistic, 25(3), 253-270.*

Boughton, K.A., Klinger, D.A. & Gierl, M. J. (2001). Effect of random rater error on parameter recovery of the generalized partial credit model and graded response model. *Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.*

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement. Fall 2004, Vol.41, No.3, Pp.261- 270.*

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory (1st ed.)*. Erlbaum Associates.

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

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Hambleton, R., & Swaminathan, H. (2013). *Item response theory: principles and applications*. Springer Science & Business Media.

Handoko, S. T., Mardiati, Y., Ismail, R., & Imawan, O. R. (2023). Employing Higher Order Thinking Skills-based Instruction in History Course: A History High School Teacher's Perspective. *AIP Conference Proceedings*, *2679*(January). https://doi.org/10.1063/5.0127631

Harsana, F. N., Retnawati, H., Dewanti, S. R., Lumenyela, R. A., Sotlikova, R., Adzima, M. F., & Septiana, A. R. (2024). Comparison of item characteristic analysis models of reading literacy test with polytomous Item Response Theory. *REID (Research and Evaluation in Education)*, *10*(2), 214–226. https://doi.org/10.21831/reid.v10i2.77852

Hortensius, L. (2012). *Advanced Measurement - Logistic regression for DIF detection*.

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

Imawan, O. R., Retnawati, H., Haryanto, & Ismail, R. (2024). Confirmatory factor analysis and differential item functioning analysis on mathematical literacy instruments for prospective Indonesian elementary school teachers. *AIP ConferenceProceeding*, 080009. https://doi.org/10.1063/5.0228174

Imawan, O. R., Retnawati, H., Haryanto, & Ismail, R. (2025). The challenges of implementing computerized adaptive testing in Indonesia. *Journal of Education and E-Learning Research*, *12*(2), 124–144. https://doi.org/10.20448/jeelr.v12i2.6677

Isgiyanto, A. (2013). Diagnosis of Student Errors Based on Polytomous Scoring Using the Partial Credit Model in Mathematics. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *15*(2), 308–325. https://doi.org/10.21831/pep.v15i2.1099

Ismail, R., Retnawati, H., Sugiman, Arovah, N. I., & Imawan, O. R. (2024). Contexts proposed by teachers in Papua for developing mathematics hots assessment instruments: A phenomenological study. *Journal of Education and E-Learning Research*, *11*(3), 548–556. https://doi.org/10.20448/jeelr.v11i3.5922

Ismail, R., Retnawati, H., Sugiman, & Imawan, O. R. (2024). Construct validity of mathematics high order thinking skills instrument with cultural context: Confirmatory factor analysis. *AIP Conference Proceeding*, 080008. https://doi.org/10.1063/5.0228143

Ismail, R., Retnawati, H., Sugiman, S., Setiawati, F. A., Imawan, O. R., & Santoso, P. H. (2024). Optimal Scale Points for Reliable Measurements: Exploring the Impact of Scale Point Variation. *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, *13*(1), 44–56. https://doi.org/10.15408/jp3i.v13i1.34173

Karimah, U., Retnawati, H., Hadiana, D., Pujiastuti, P., & Yusron, E. (2021). The characteristics of chemistry test items on nationally-standardized school examination in Yogyakarta City. *REID (Research and Evaluation in Education)*, *7*(1), 1–12. https://doi.org/10.21831/reid.v7i1.31297

Kartianom, K., & Mardapi, D. (2018). The utilization of junior high school mathematics national examination data: A conceptual error diagnosis. *REID (Research and Evaluation in Education)*, *3*(2), 163–173. https://doi.org/10.21831/reid.v3i2.18120

Kartowagiran, B., Mardapi, D., Purnama, D. N., & Kriswantoro, K. (2019). Parallel tests viewed from the arrangement of item numbers and alternative answers. *REID (Research and Evaluation in Education)*, *5*(2), 169–182. https://doi.org/10.21831/reid.v5i2.23721

Kusumawati, M., & Hadi, S. (2018). *An analysis of multiple choice questions (MCQs): Item and test statistics from mathematics assessments in senior high school*. *4*(1), 70–78. http://journal.uny.ac.id/index.php/reid

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Lei Chang. (1994). A Psychometric evaluation of 4-poin and 6-point Likert-type scales in relatiopn to reliability and validity. [Versi elektronik]. *Apllied Psychological Measurement, 18, 3, 205-215*.

Lin, C. J. (2008). Comparisons between classical test theory and butir response theory in automated assembly of parallel test form. *The Journal of Technology, Learning, and Assessment. 6(8), 1-42.*

McGrath, J. M., & Brandon, D. (2018). What Constitutes a Well-Designed Pilot Study? *Advances in Neonatal Care*, *18*(4), 243–245. https://doi.org/10.1097/ANC.0000000000000535

Messick, S. J. (1995). Validity of Psychological Assesment: Validation of Infrences from Persons Responses and Performance as Scientific Inquiry Into Score Meaning. *American Psychologist, 50 (9), 741-749.*

Otaya, L. G., Kartowagiran, B., Retnawati, H., & Mustakim, S. S. (2020). Estimating the ability of pre-service and in-service Teacher Profession Education (TPE) participants using Item Response Theory. *REID (Research and Evaluation in Education)*, *6*(2), 160–173. https://doi.org/10.21831/reid.v6i2.36043

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *REID (Research and Evaluation in Education)*, *9*(1), 86–117. https://doi.org/10.21831/reid.v9i1.63230

Ploutz-Snyder, L., Bloomfield, S., & Smith, S. M. (2014). Fundamental Principles of Small Sample Size Research. *Frontiers in Physiology*, *5*, 413.

Susongko, P. (2010). Perbandingan Keefektifan Bentuk Tes Uraian Dan Testlet Dengan Penerapan Graded Response Model (GRM). *Jurnal Penelitian Dan Evaluasi Pendidikan*.

Retnawati, H. (2014). *Teori respon butir dan penerapannya*. Parama Publishing.

Robitzsch, A. (2025). sirt: Supplementary Item Response Theory Models. In *CRAN: Contributed Packages* (pp. 12–80). https://doi.org/10.32614/CRAN.package.sirt

Santoso, P. H., Setiawati, F. A., Ismail, R., & Suhariyono, S. (2023). Comparing IRT properties among different category numbers: a case from attitudinal measurement on physics education research. *Discover Psychology*, *3*(1). https://doi.org/10.1007/s44202-023-00101-6

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., Graeff, A. de, Groenvold, M., Gundy, C., & Koller, M. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales Scott. *Journal of Clinical Epidemiology*, *62*(3), 288–295.

Setiawan, A., Kassymova, G. K., Mbazumutima, V., & Agustyani, A. R. D. (2024). Differential Item Functioning of the region-based national examination equipment. *REID (Research and Evaluation in Education)*, *10*(1), 99–113. https://doi.org/10.21831/reid.v10i1.73270

Sheppard, R., dkk. (2006). *Differential Item Functioning by Sex and Race in the Hogan Personality Inventory*.

Strout, W. F. (1990). A New Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation. *Psychometrika*, *55*(2), 293–325. https://doi.org/10.1007/BF02295289

Sumin, S., Sukmawati, F., & Nurdin, N. (2022). Gender differential item functioning on the Kentucky Inventory of Mindfulness Skills instrument using logistic regression. *REID (Research and Evaluation in Education)*, *8*(1), 55–66. https://doi.org/10.21831/reid.v8i1.50809

Sumintono, B. & Widhiarso, W. (2013). *Aplikasi Model Rasch: untuk penelitian ilmu sosial. Edisi 1.* Trim Komunikata Publishing House.

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Tang, K. L. (1996). Polytomous item response theory (IRT) models and their aplications in large-scale testing program: Review of literature. *Educational Testing Science. Princeton, NJ. RM-96-8 TOEFL Monograph Series.*

Tognolini, J., & Davidson, M. (2003). How do we operationalise what we value? Some technical chalenges in assessing higher order thinking skills. *Makalah Disajikan Dalam the Natinaonal Roundtable on Assessment Conference Pada Bulan Juli 2003 Di Darwin, Australia.*

Wardani, R. E. A., Prihatni, Y., Negeri, S., & Jl Jogja-Solo Km, K. (2018). *Developing assessment model for bandel attitudes based on the teachings of Ki Hadjar Dewantara*. *4*(2), 117–125. http://journal.uny.ac.id/index.php/reid

Wasis. (2011). Model Penskoran Partial Credit Pada Butir Multiple True-False Bidang Fisika. *Jurnal Penelitian Dan Evaluasi Pendidikan*.

Wu, B. C. (2003). *Scoring multiple true-false butirs: A comparison of summed scores and response pattern scores at butir and test level. Research report*. Educational Resources International Center (ERIC).

Yim, L. W. K., Lye, C. Y., & Koh, P. W. (2024). A psychometric evaluation of an item bank for an English reading comprehension tool using Rasch analysis. *REID (Research and Evaluation in Education)*, *10*(1), 18–34. https://doi.org/10.21831/reid.v10i1.65284

Yudiana, W., Triwahyuni, A., & Susanto, H. (2023). Multidimensional Rasch Analysis of Gender Differences in Tes Intelegensi Kolektif Indonesia–Tinggi (TIKI-T). *JP3I (Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia)*, *12*(1), 1–16. https://doi.org/10.15408/jp3i.v12i1.20417

Zhang, J. (2006). Conditional Covariance Theory and Detect for Polytomous Items. *Psychometrika*, *72*(1), 69–91. https://doi.org/10.1007/s11336-004-1257-7

This is an open access article under CC-BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)