AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 14, No. 4, 2025, 1182–1210

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

GPCM ANALYSIS OF CONSTRUCTED-RESPONSE ASSESSMENT (CRA): MEASURING ABSTRACTION VIA ANALYTIC CODING

Dwi Rismi Ocy[1*], Wardani Rahayu[2], Iva Sarifah[3], Awaluddin Tjalla[4]

[1,2,3,4] Universitas Negeri Jakarta, Jakarta, Indonesia
**Corresponding author. Jl. Kota Piring, 29123, Tanjung Pinang, Indonesia.*
E-mail: *dwirismiocy@gmail.com*[*1]

Received 04 June 2025; Received in revised form 19 November 2025; Accepted 08 December 2025

Abstract

Traditional multiple-choice tests frequently fail to capture the subtle cognitive ability of mathematical abstraction in subjects like trigonometric ratio. The purpose of this study was to create and evaluate a performance-based evaluation tool to gauge students' proficiency with mathematical abstraction in the context of trigonometric ratio. Three constructed-response items made up the instrument, and each one was assessed using a unique analytical rubric that was in line with revised Bloom's taxonomy and abstraction processes. Following person-fit screening, 700 replies from 913 tenth-grade students in six Indonesian high schools were kept. Strong item fit, suitable threshold values, and high discrimination indices were found via Generalized Partial Credit Model (GPCM) analysis, demonstrating the instrument's efficacy in distinguishing between different levels of abstraction. Unidimensionality was confirmed by exploratory factor analysis, which had an excellent model fit and explained 41.0% of the total variance (RMSEA = 0.0301; TLI = 0.9775). Item difficulty levels matched the sample's ability distribution quite well, according to person-item mappings. The construct validity of the analytical rubric was further reinforced by high factor loadings and communalities. Results show that the created tool is theoretically and psychometrically solid, providing a strong means of evaluating Mathematical Abstraction in HOTS. The study demonstrates how GPCM-based analytical scoring may be used to capture complex cognitive performance and guide teaching strategies in accordance with the Merdeka Curriculum.

Keywords: Analytic rubric; constructed-response assessment; EFA; GPCM; IRT; mathematical abstraction.

Abstrak

Tes pilihan ganda tradisional seringkali gagal menangkap kemampuan kognitif yang subtil dari abstraksi matematis dalam mata pelajaran seperti perbandingan trigonometri. Tujuan dari penelitian ini adalah untuk membuat dan mengevaluasi alat evaluasi berbasis kinerja untuk mengukur kemahiran siswa dalam abstraksi matematis dalam konteks perbandingan trigonometri. Instrumen ini terdiri dari tiga item respons terbuka, dan masing-masing dinilai menggunakan rubrik analitik unik yang selaras dengan taksonomi Bloom yang direvisi dan proses abstraksi. Setelah penyaringan person-fit, 700 jawaban dari 913 siswa kelas sepuluh di enam sekolah menengah atas di Indonesia dipertahankan. Analisis Generalized Partial Credit Model (GPCM) mengungkapkan kecocokan item yang kuat, nilai ambang batas yang sesuai, dan indeks diskriminasi yang tinggi, yang menunjukkan efikasi instrumen dalam membedakan antara tingkat abstraksi yang berbeda. Unidimensionalitas dikonfirmasi oleh analisis faktor eksploratori, yang memiliki kecocokan model yang sangat baik dan menjelaskan 41,0% dari total varians (RMSEA = 0,0301; TLI = 0,9775). Tingkat kesulitan item sangat cocok dengan distribusi kemampuan sampel, menurut pemetaan person-item. Validitas konstruk rubrik analitik lebih lanjut diperkuat oleh pemuatan faktor dan komunalitas yang tinggi. Hasil penelitian menunjukkan bahwa alat yang dibuat ini solid secara teoritis dan psikometris, menyediakan sarana yang kuat untuk mengevaluasi Abstraksi Matematis dalam HOTS. Studi ini menunjukkan bagaimana penskoran analitik berbasis GPCM dapat digunakan untuk menangkap kinerja kognitif yang kompleks dan memandu strategi pengajaran sesuai dengan Kurikulum Merdeka.

***Kata kunci**: Abstraksi matematis; asesmen respons terkonstruksi; EFA; GPCM; rubrik analitik; MIRT*

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

INTRODUCTION

Mathematics, being a diverse discipline, requires not just numerical computation but also the development of abstract thinking skills, which are essential for a profound understanding of the subject. When addressing complex topics such as trigonometric ratio, evaluating students' ability to abstract mathematical concepts is crucial. Mathematical abstraction enables learners to develop deeper insights, establish structural links among various concepts, and visualize challenges (Syarifudin et al., 2021). While significant, existing assessment instruments often reduce this cognitive intricacy to multiple-choice formats, failing to capture the depth of students' thought processes (Hawai, 2021).

This reductionist approach to evaluation techniques leads to significant consequences. The limitations of multiple-choice items in accurately representing students' mental models can lead to the neglect of more sophisticated cognitive skills such as mental transformation, generalization, and reasoning (Attali et al., 2016; Kuo et al., 2016). These limitations are particularly evident in areas such as trigonometric ratio, where learners must internalize, represent, and apply abstract relationships within appropriate contexts (Sukmawan et al., 2022). Assessment instruments should evolve to more accurately capture the nuances of students' abstract thinking.

Constructed-response (CR) items offer a broader platform for students to articulate their thoughts through symbolic, visual, and spoken methods (Edimuslim, 2022). CR assignments push students to engage deeply by analyzing, justifying, and creatively interacting with mathematical scenarios, thereby fostering more authentic and contextually relevant learning experiences (Khairunnisa et al., 2021). Furthermore, CR has demonstrated significant alignment with educational innovations that foster critical thinking and comprehensive cognitive engagement (Faisal et al., 2020; Ocy et al., 2023; Yanti & Wijaya, 2023).

The critique of multiple-choice tests remains valid and persistent. When faced with unfamiliar or non-routine problems, these tools often restrict students' ability to express complex understanding (Angraini, 2018; Erni et al., 2022). This critique emphasizes the necessity for assessment reforms that promote deeper thinking among students and consider the real-world challenges of solving mathematical problems.

A study conducted by Syarifudin et al. (2021) at MAN 1 Tasikmalaya indicates that students who possess higher abstraction skills demonstrate an enhanced capacity to generalize mathematical concepts and translate problems into appropriate symbolic representations. Consequently, improving mathematical education requires implementing concrete strategies to evaluate abstract reasoning and offering students opportunities to engage in creative problem-solving during assessments.

As a result, analytic coding rubrics have progressively emerged as more transparent alternatives to tradetional scoring systems. Analytic rubrics allow for a more detailed evaluation of specific aspects of student performance compared to holistic or numerical scoring, thereby facilitating clearer and more actionable feedback (Karakuş & Ocak, 2022). Analytic rubrics provide valuable insights not only into outcomes but also into the processes students engage in, thereby enhancing the alignment between assessment and

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

instructional objectives in the evaluation of mathematical abstraction (Amelia et al., 2024).

Despite their potential, the incurporation of constructed-response items, analytic rubrics, and psychometric validation through models like the Generalized Partial Credit Model (GPCM) continues to face significant theoretical and methodological challenges. While each of these components contributes to enhancing assessment quality, their collective application—particularly in specialized areas such as trigonometry—has not been extensively explored. GPCM is especially suitable for polytomous scoring systems as it accommodates variations in performance levels and provides a more nuanced understanding of student learning trajectories.

This study addresses the existing gap by innovatively integrating constructed-response items developed under the mathematical abstraction theory with the revised Bloom's taxonomy, resulting in an analytic rubric designed for assessing mathematical abstraction. This framework enhances the articulation of anticipated student performance and aligns measurement techniques with genuine cognitive demands in trigonometry (Navas-López, 2024; Ngu & Phan, 2020). When effectively implemented, analytical rubrics provide educators and learners with a structured approach to addressing challenges in mathematics.

Currently, there is a notable lack of psychometric validation for these tests utilizing GPCM within the existing literature. Future investigations must address this discrepancy to enhance the instructional relevance, interpretability, and reliability of mathematical assessments. Therefore, this study aims to examine the psychometric validity of a constructed-response assessment focused on trigonometric ratio, evaluated through an analytic rubric, utilizing the Generalized Partial Credit Model (GPCM). This investigation focuses on item fit, threshold ordering, dependability, and indicators of construct validity within the IRT framework.

METHODS

Research Design

This study employed a quantitative research approach, focusing on the psychometric validation of a performance-based assessment instrument designed to measure students' mathematical abstraction abilities in the context of trigonometric ratio. The analysis utilized the Generalized Partial Credit Model (GPCM) to examine the functioning of the constructed-response items and the analytic scoring rubric. The study concentrated on evaluating item fit, category functioning, reliability, and the overall validity of the scoring framework, aiming to ensure that the instrument accurately captures the targeted construct of mathematical abstraction.

Participants

This study involved six senior high schools located in Sukabumi and Tanjungpinang, Indonesia. The selected institutions comprised both public and private schools, each exhibiting diverse academic profiles. The participating schools comprised SMAN 1 Sukaraja, SMAN 1 Cireunghas, SMAN 1 Tanjungpinang, SMAN 2 Tanjungpinang, SMAS Maitreyawira, and SMAS Santa Maria. By prioritizing advanced cognitive skills in their instructional and evaluative approaches, these institutions have all embraced the Merdeka Curriculum. The qualities outlined were essential as the

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

instrument created in this study aimed to assess students' mathematical abstraction ability—an advanced cognitive construct—utilizing indicators that correspond with abstraction processes, while ensuring that item difficulty levels were calibrated to represent varying degrees of higher-order thinking skills. Schools that intentionally integrated higher-order thinking skills in their teaching methods and evaluations were selected to ensure a cohesive relationship between the educational environment and the requirements of the assessment tool.

The initial assessment drew a total of 913 students from the tenth grade. A total of 700 student responses were retained for the final psychometric analysis following a comprehensive data screening and model fit assessment using the Generalized Partial Credit Model (GPCM). The inadequate fit of the data model, indicated by person-level infit and outfit mean square statistics exceeding acceptable thresholds, resulted in the exclusion of a total of 213 responses. This enhancement ensured that the final dataset met the GPCM assumptions, thereby reinforcing the validity of the estimated item parameters and individual ability measures.

Data Analysis

The polytomous scoring data were analyzed using the Generalized Partial Credit Model (GPCM) via the *mirt* package in R. Prior to IRT modeling, item response data were screened for missing responses, response sets, and extreme scores. A total of 700 student responses met the criteria for person-fit based on outfit and infit mean square statistics and were retained for further analysis.

The first analytical step involved testing the unidimensionality assumption using Exploratory Factor Analysis (EFA) with the EFA.dimensions package in R, which is specifically designed for dimensionality assessment in item response data. The EFA results supported a unidimensional structure, satisfying a key prerequisite for subsequent IRT modeling (de Ayala, 2009; Prihono et al., 2022; Watson, 2017).

Next, model-data fit comparisons were conducted to determine the most appropriate IRT model for the data. Three polytomous IRT models—Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM), and Graded Response Model (GRM)—were compared based on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), log-likelihood values, and M2 statistics (Bürkner et al., 2018; Wind, 2022). The GPCM demonstrated superior model fit across these indices and was selected for final analysis.

Following model selection, GPCM analysis was performed to estimate item parameters, including step difficulties (thresholds) and item discrimination indices (Bürkner et al., 2018; Wind, 2022; Zhou & Huggins-Manley, 2020). Item fit statistics were evaluated to ensure model adequacy. EAP reliability coefficients were computed to assess the internal consistency of the instrument. Finally, person-item maps (Wright maps) were generated to examine the alignment between item difficulty and student ability levels, providing further support for the construct validity and measurement precision of the instrument.

RESULTS AND DISCUSSION

Instrument Development

The assessment instrument consisted of three constructed-response items designed to measure students' mathematical abstraction ability in the context of trigonometric ratios. The

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

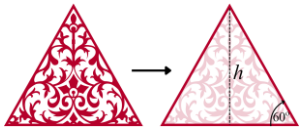
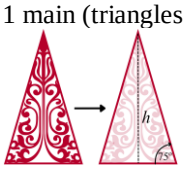
DOI: https://doi.org/10.24127/ajpm.v14i4.12976

items were developed based on specific indicators of mathematical abstraction ability and aligned with revised Bloom's taxonomy, aiming to capture cognitive processes ranging from conceptual understanding to abstract generalization.

To score student responses, a custom analytic rubric was developed. Each item's rubric consisted of multiple indicators, each reflecting specific sub-levels within mathematical abstraction (i.e., perceptual abstraction, internalization, interiorization, second level of interiorization). The rubric levels varied by indicator, ranging from 0–3 or 0–4 points, depending on the complexity of the expected response.

To measure students' mathematical abstraction abilities within real-world problem-solving contexts, this research utilized a Constructed-Response Assessment (CRA). CRA items are specifically designed to encourage students to demonstrate a deep understanding and application of mathematical concepts, moving beyond simple multiple-choice answers. This particular item aims to assess students at the C5 (Evaluating) cognitive level, where they must judge and decide on the best solution based on mathematical criteria. Specifically, the question promotes the Interiorization level of abstraction, as students are expected to mentally manipulate trigonometric concepts, integrate various knowledge points (e.g., calculating triangle bases, motif units, space efficiency), and apply them in constructing a mathematical model to solve the given design problem. One of the CRA items used in this study, focusing on Trigonometric Ratios, is presented in Table 1.

Table 1. Item Card Sample Constructed-Response Item (Trigonometric Ratios)

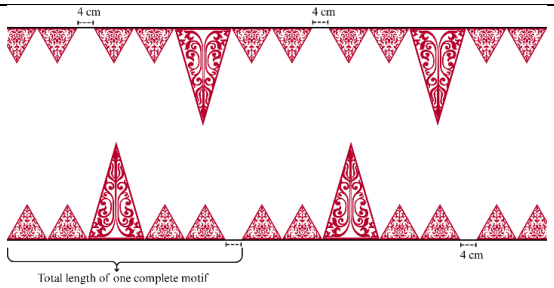
Item Card					
Identification			**Specification**		
Subject	:	Mathematics	Cognitive Level	:	**C5 (Evaluating)**
Grade	:	10^{th} Grade	Basic Competency	:	Solving contextual problems related to trigonometric ratios
Semester	:	1^{st} Semester	Item Type	:	Constructed-Response (Essay)
Topic	:	Trigonometric Ratios	Indicator	:	Students are able to evaluate the spatial efficiency of two different "pucuk rebung" motif designs using trigonometric ratios.
Abstraction Level	:	**Interiorization**			
Stimulus					
Ani received an order to create a 3 × 1.5 meter batik cloth with a "pucuk rebung" motif. She is considering two design alternatives: **Design A (Previously Used by Ani)** One complete motif consists of: • 4 side motifs "pucuk rebung" (triangle with 60° angle, height (t) = $8\sqrt{3}$ cm)  • 1 main (triangles with 75° angle, height (t) = $15(\sqrt{3})$ cm  • An additional 4 cm of spacing is required on both ends.					

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 14, No. 4, 2025, 1182–1210
ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Item Card

Stimulus



Total length of one complete motif: (4 × base of side triangle) + base of main triangle + 4 cm

Design B (New Design)

Each motif consists of:

- 2 main motifs (same height and angle as in A)
- 2 side motifs (same height and angle as in A)

No additional spacing is needed.

Question (Item Number 1)

1. Using the provided information about both "pucuk rebung" motif designs, complete the following three tasks in order. Show all your calculation steps clearly. ($\sqrt{3} \approx 1.732$).
   a. Calculate how many complete motifs can fit along the 3-meter lower edge of the cloth for each design.
   b. Evaluate the space efficiency of both designs.
   c. Decide which design is better and justify your answer using mathematical reasoning.

Solution

I. Pre-computation of Base Lengths

To calculate the number of motifs, the base length of each triangle is required. It is assumed that the given angle (75° or 60°) is a base angle of the isosceles triangle forming the "pucuk rebung" motif, and that the height forms a right-angled triangle where the given angle is one of the acute angles. In such a right-angled triangle, $\tan\theta = \frac{Opposite}{Adjacent}$, where the height is the opposite side and half of the base is the adjacent side.

1. Main "Pucuk Rebung" Triangle (75° angle, height $= 15\sqrt{3}$ cm)

   Let b_M be base of the main triangle.
   Height $h_m = 15\sqrt{3}$ cm.
   Assuming 75° is a base angle:
   $\tan 75° = \frac{h_M}{b_M/2}$

   We know that $\tan 75° = \tan(45° + 30°) = \frac{\tan 45° + \tan 30°}{1 - \tan 45° \tan 30°} = \frac{1+\frac{1}{\sqrt{3}}}{1-\frac{1}{\sqrt{3}}} = \frac{\sqrt{3}+1}{\sqrt{3}-1} = \frac{(\sqrt{3}+1)^2}{3-1} = 2 + \sqrt{3}$

   So, $\frac{b_M}{2} = \frac{h_M}{\tan 75°} = \frac{15\sqrt{3}}{2+\sqrt{3}}$
   To rationalize the denominator, multiply by the conjugate $(2 - \sqrt{3})$:
   $\frac{b_M}{2} = \frac{15\sqrt{3}}{2+\sqrt{3}} \times \frac{(2-\sqrt{3})}{(2-\sqrt{3})} = 30\sqrt{3} + 45.$
   Therefore,
   $b_M = 2(30\sqrt{3} - 45) = 60\sqrt{3} - 90$ cm. (Using $\sqrt{3} \approx 1.732$)
   $b_M = 60(1.732) - 90 = 13.92$ cm.

2. Side Motif Triangle (60° angle, height $= 8\sqrt{3}$ cm)
   Let b_S be the base of the side triangle. Height $h_S = 8\sqrt{3}$ cm.
   Assuming 60° is a base angle:
   $\tan 60° = \frac{h_S}{b_S/2} \rightarrow \sqrt{3} = \frac{8\sqrt{3}}{b_S/2} \rightarrow \frac{b_S}{2} = \frac{8\sqrt{3}}{\sqrt{3}} = 8$ cm.
   Therefore, $b_S = 16$ cm.

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 14, No. 4, 2025, 1182–1210
ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Item Card

Solution

II. Calculate Motif Lengths and Number of Motifs

Total length of the cloth = 3 meters = 300 cm.

Design A:

One complete motif unit consists of: 1 main "pucuk rebung" 4 side motifs An additional 4 cm of spacing is required on both ends. This typically means 4 cm at the beginning of the entire layout and 4 cm at the end. So, 2×4=8 cm total fixed spacing.

Total length of one complete motif (as described in the problem statement for Design A's unit):

$L_{A,unit} = (4 \times base\ of\ side\ triangle) + base\ of\ main\ triangle + 4\ cm.$

Given "An additional 4 cm of spacing is required on both ends" in the general description for Design A, and then "Total length of one complete motif: (4 × base of side triangle) + base of main triangle + 4 cm", it's most logical to interpret the latter "4 cm" as *part of the repeating unit's length*, which already includes the internal spacing logic, while the former "on both ends" applies to the overall arrangement, meaning the effective length for placing these repeating units is $300 - (2 \times 4) = 292$ cm. This implies the "4 cm" in the unit length is an internal spacing.

Let's proceed with this interpretation:

$L_{A,unit} = (4 \times b_S) + b_M + 4\ cm$

$L_{A,unit} = (4 \times 16) + (60\sqrt{3} - 90) + 4\ cm$

$L_{A,unit} = 64 + 60\sqrt{3} - 90 + 4$

$L_{A,unit} = 60\sqrt{3} - 22 \approx 81.92$ cm

Effective length for motifs in Design A (after overall fixed spacing):

Effective length $= 300 - (2 \times 4)cm = 300 - 8\ cm = 292\ cm.$

Number of complete motifs for Design A:

Number of motifs $= \frac{Effective\ length}{Length\ of\ one\ unit} = \frac{292}{81.92} = 3.564 =$ **3 *complete motifs.***

Design B:

Each motif consists of:

2 main motifs

2 side motifs

No additional spacing is needed.

Length of one repeatable unit in Design B:

Number of motifs $= \left\lfloor \frac{Total\ length}{Length\ of\ one\ unit} \right\rfloor = \left\lfloor \frac{300}{59.84} \right\rfloor = \lfloor 5.013 \rfloor =$ **5 *complete motifs.***

III. Evaluate Space Efficiency and Justify

Space Efficiency Evaluation:

Design A:

Total length covered by 3 motifs (including their internal 4cm spacing) = $3 \times 81.92 = 245.76$ cm.

Total length occupied including initial/final 4 cm spacing = $245.76 + 8 = 253.76$ cm.

Unused space = $300 - 253.76 = 46.24$ cm.

Percentage of total length used by motifs (excluding fixed outer spacing):

$\frac{245.76}{300} \times 100\% \approx 81.92\%$

Percentage of total length used by motifs and their required spacing: $\frac{253.76}{300} \times 100\% \approx 84.59\%$.

Design B:

Total length covered by 5 motifs = $5 \times 59.84 = 299.2$ cm.

Unused space = $300 - 299.2 = 0.8$ cm.

Percentage of total length used by motifs:

$\frac{299.2}{300} \times 100\% = 99.73\%$.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Item Card
Solution
Decision and Justification: Design B allows for significantly more complete motifs (5 motifs) compared to Design A (3 motifs) along the 3-meter lower edge of the cloth. Furthermore, Design B is vastly more space-efficient: • Design A covers approximately 81.92% of the cloth with actual motif elements (or 84.59% including its internal and external required spacing), leaving a substantial unused space of 46.24 cm. • Design B covers approximately 99.73% of the cloth with motif elements, leaving a minimal unused space of only 0.8 cm. Mathematically, Design B maximizes the utilization of the cloth's length for the "pucuk rebung" pattern. This leads to a fuller, more dense design and significantly. less wasted space on the cloth, making it the more efficient and economically sound choice for Ani

This item was developed due to its inherent characteristic of requiring students not just to calculate, but also to apply trigonometric ratios in complex contextual situations and perform a spatial evaluation. The open-ended nature of this question allows students to exhibit a range of responses, from fundamental conceptual errors to sophisticated mathematical modeling and data-driven justifications. This variation in responses is crucial for a deeper abstraction analysis, as it elicits diverse cognitive and abstraction levels that can be identified. The scoring for this item was conducted using an analytic coding rubric, which will be further elaborated in Table 2, ensuring consistency and objectivity in the *analytic coding* assessment that is the focus of this research.

To ensure objectivity and consistency in scoring student responses to the CRA item, particularly in identifying mathematical abstraction levels, an analytic scoring rubric was employed. This rubric is designed to break down the task into measurable steps and provide clear criteria for each performance level. This approach allows for a detailed assessment, not only of the final answer but also of the mathematical thought process demonstrated by the students. The analytic rubric used for the Trigonometric Ratios item, focusing on *analytic coding* to measure abstraction, is presented in Table 2.

Table 2. Corresponding Analytic Coding Rubric

Evaluation Step	Score 0	Score 1	Score 2	Score 3	Score 4
Calculating the Base Length of the Main "Pucuk Rebung" Motif (Main Motif)	Unable to calculate the base length of the main motif or the calculation is completely incorrect.	Calculates the base length of the main motif but with fundamental trigonometric conceptual errors (e.g., incorrect function used, incorrect angle placement, or significant algebraic errors).	Calculates the base length of the main motif with correct trigonometric concepts, but there are minor calculation errors or imprecise rounding.	Calculates the base length of the main motif correctly and accurately, demonstrating a strong understanding of trigonometric concepts	

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Evaluation Step	Score 0	Score 1	Score 2	Score 3	Score 4
Calculating the Base Length of the Side Motif	Unable to calculate the base length of the side motif or the calculation is completely incorrect.	Calculates the base length of the side motif but with fundamental trigonometric conceptual errors (e.g., incorrect function used, incorrect angle placement, or significant algebraic errors).	Calculates the base length of the side motif with correct trigonometric concepts, but there are minor calculation errors or imprecise rounding.	Calculates the base length of the side motif correctly and accurately, demonstrating a strong understanding of trigonometric concepts.	
Calculating the Number of Motifs for Design A	Unable to calculate the number of motifs for Design A or the calculation is completely incorrect.	Calculates the number of motifs for Design A but there are errors in calculating the motif unit length or handling the spacing.	Calculates the number of motifs for Design A with correct motif unit length and spacing handling, but there are rounding errors or imprecise interpretation of the floor function ($\lfloor\rfloor$).	Calculates the number of motifs for Design A correctly and accurately, including proper spacing handling and the use of the floor function.	
Calculating the Number of Motifs for Design B	Unable to calculate the number of motifs for Design B or the calculation is completely incorrect.	Calculates the number of motifs for Design B but there are errors in calculating the motif unit length.	Calculates the number of motifs for Design B with correct motif unit length, but there are rounding errors or imprecise interpretation of the floor function ($\lfloor\rfloor$).	Calculates the number of motifs for Design B correctly and accurately, including the proper use of the floor function.	
Evaluating Space Efficiency (Percentage / Remaining Space Calculation)	Unable to evaluate space efficiency or performs irrelevant calculations.	Evaluates space efficiency but the percentage or remaining space calculations are completely incorrect or inconsistent.	Evaluates space efficiency by performing largely correct percentage or remaining space calculations, but there are minor errors or it's incomplete.	Evaluates space efficiency with correct and accurate percentage and remaining space calculations for both designs.	
Drawing a Conclusion and Justification	Unable to draw a conclusion, or the conclusion is irrelevant/unsupported.	Draws a conclusion but the justification is absent, irrelevant, or very weak.	Draws a correct conclusion, but the justification is not strong enough or not entirely based on the calculated mathematical data.	Draws a correct conclusion and provides logical justification, but the mathematical explanation is not deep enough or lacks structure.	Draws a correct conclusion and provides strong, logical, in-depth mathematical justification, fully supported by the space efficiency calculation data, and presented clearly.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

This analytic scoring rubric serves as a key instrument in the data analysis, enabling researchers to systematically categorize and score each component of the students' answers. The clear criteria at each step, including the score interpretations for each abstraction level, facilitate the identification of nuances in students' understanding and application of trigonometric concepts. The data collected through scoring with this rubric will then be analyzed using the Generalized Partial Credit Model (GPCM), a suitable Rasch model for such partially scored data, to measure students' mathematical abstraction abilities more accurately and comprehensively.

Student responses were rated using this rubric and then converted into polytomous numerical data. Each rubric level was treated as an ordered response category, allowing the data to be modeled using the Generalized Partial Credit Model (GPCM). In the context of psychometric modeling, each indicator within the rubric was treated as a separate item, resulting in a fine-grained item-level analysis rather than scoring the constructed-response item as a single holistic score.

This approach enabled detailed analysis of student performance across different dimensions of abstraction and supported the examination of step-level thresholds in item functioning.

Exploratory Factor Analysis Results

To assess the unidimensionality assumption—a key requirement for applying the IRT Models—an exploratory factor analysis (EFA) was conducted on all constructed-response items scored using analytic coding. Each item (e.g., ITEM1I, ITEM2VI) represents a specific scoring criterion within the respective constructed-response task. For instance, ITEM1I refers to Item 1, Criterion I, while ITEM2VI corresponds to Item 2, Criterion VI. The analysis aimed to verify whether all scoring criteria across the three constructed-response items measure a single latent construct: mathematical abstraction in the context of trigonometric ratio.

Table 3 presents the factor loadings of each scoring criterion on the first extracted factor (MR1), alongside communalities (h^2), uniqueness values (u^2), and complexity indices. A strong loading on a single factor and low complexity suggest that the analytic rubric captures a coherent, unidimensional construct.

Table 3. Factor Analysis Results (Unidimensionality Assumption)

Item	Loading MR1	h^2 (Communality)	u^2 (Uniqueness)	Complexity
ITEM1I	0.479	0.229	0.771	1
ITEM1II	0.729	0.531	0.469	1
ITEM1III	0.689	0.475	0.525	1
ITEM1IV	0.724	0.524	0.476	1
ITEM1V	0.762	0.580	0.420	1
ITEM1VI	0.794	0.631	0.369	1
ITEM2I	0.331	0.110	0.890	1
ITEM2II	0.553	0.305	0.695	1
ITEM2III	0.703	0.494	0.506	1
ITEM2IV	0.645	0.416	0.584	1
ITEM2V	0.743	0.552	0.448	1
ITEM2VI	0.668	0.446	0.554	1
ITEM2VII	0.555	0.308	0.692	1

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Item	Loading MR1	h^2 (Communality)	u^2 (Uniqueness)	Complexity
ITEM2VIII	0.761	0.580	0.420	1
ITEM2IX	0.563	0.317	0.683	1
ITEM2X	0.727	0.529	0.471	1
ITEM2XI	0.557	0.310	0.690	1
ITEM2XII	0.681	0.464	0.536	1
ITEM3I	0.526	0.277	0.723	1
ITEM3II	0.527	0.278	0.722	1
ITEM3III	0.562	0.316	0.684	1
ITEM3IV	0.594	0.353	0.647	1

The results in Table 3 show that the majority of the analytic scoring criteria exhibit moderate to high loadings on the first factor, with loadings ranging from 0.331 to 0.794 (Gos et al., 2020). All items have complexity values of 1.0, indicating they primarily load on a single factor, which supports the unidimensionality assumption (Shrestha, 2021; Watkins, 2018). While ITEM2I showed a relatively low loading (0.331), most items—including those from Items 1, 2, and 3—demonstrated substantial contributions to the latent construct. These findings suggest that the analytic coding rubric effectively captures a unified dimension of mathematical abstraction across diverse criteria and response structures, providing initial evidence of construct validity for the rubric in the context of constructed-response assessment.

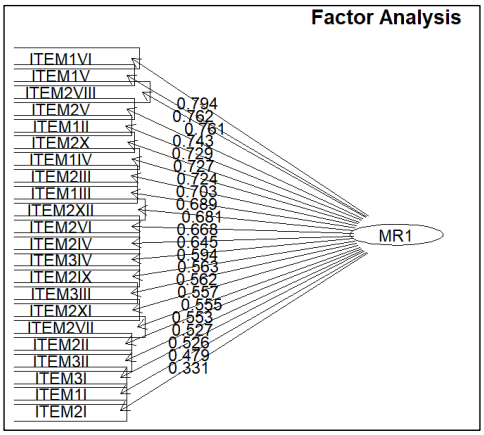


Figure 1. Factor analysis

Figure 1 visually illustrates the factor loadings of each analytic scoring criterion on the single extracted factor, MR1. Consistent with the numerical results in Table 3, most items show moderate to high loadings, reinforcing the strong relationship between these items and the underlying construct of mathematical abstraction. Although ITEM2I has the lowest loading (0.331), it still contributes to the factor without compromising the overall unidimensionality. This graphical representation further supports the construct validity of the analytic rubric by confirming that the scoring criteria collectively measure a unified latent dimension.

To further evaluate the dimensional structure of the analytic scoring data and strengthen the evidence for unidimensionality, several global fit indices were examined based on the results of the exploratory factor analysis. These indices offer complementary information to the item-level loadings in Table 3 and help determine how well the one-factor model fits the observed data. Table 4 summarizes the key model fit statistics, including variance explained, residual-based indices, and fit coefficients such as RMSEA and TLI. This analysis provides essential evidence regarding the appropriateness of applying a unidimensional item response model such as the GRM, PCM, and GPCM to the constructed-response items scored using the analytic rubric.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Table 4. Summary statistics of factor analysis

Statistic	Value
Number of Extracted Factors	1
SS Loadings	9.025
Proportion of Variance Explained by Factor 1	41.0%
RMSR (Root Mean Square Residual)	0.027
df-corrected RMSR	0.028
Empirical Chi-Square (df = 209), p-value	231.674, p = 0.135
Likelihood Chi-Square, p-value	342.024, p < 0.0000000174
TLI (Tucker Lewis Index)	0.9775
RMSEA (90% Confidence Interval)	0.0301 (0.0243 – 0.0358)
BIC	-1027.152
Model Fit (based on off-diagonal values)	0.996
Correlation of factor scores with factor	0.972
R^2 of scores with factor	0.945
Minimum possible correlation of factor scores	0.890

As shown in Table 4, the single-factor model explains 41.0% of the total variance, with SS loadings of 9.025, which is acceptable given the complexity and multidimensional potential of constructed-response items (Watkins, 2018; Yao & Schwarz, 2006). The RMSR (0.027) and df-corrected RMSR (0.028) indicate a good residual fit (Flora & Flake, 2017; Schneider et al., 2020). Importantly, the RMSEA value of 0.0301, along with a tight 90% confidence interval (0.0243 – 0.0358), and a high Tucker Lewis Index (TLI = 0.9775) suggest a strong overall model fit (Lorenzo-Seva & Ferrando, 2023; Shrestha, 2021). Moreover, the empirical chi-square test yielded a non-significant result (p = 0.135), which supports the null hypothesis that the one-factor model fits the data well ((Flora & Flake, 2017); (Gibson Jr. et al., 2020); (Watkins, 2018)).

Additional indicators, such as the high model fit based on off-diagonal values (0.996), strong factor-score correlations (r = 0.972), and high explained variance in factor scores (R^2 = 0.945), further confirm the coherence of the unidimensional model ((de Ayala, 2009); (Gibson Jr. et al., 2020)). Collectively, these results reinforce the validity of using an analytic coding rubric to capture a single latent construct—mathematical abstraction—across all three constructed-response items in this assessment.

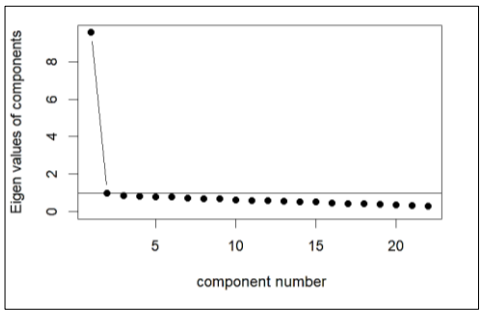


Figure 2. Scree plot

Figure 2 presents the scree plot, which clearly shows a steep decline after the first component, with the first eigenvalue substantially larger than the rest. This visual pattern aligns with the statistical findings in Table 4, where the single-factor model accounts for 41.0% of the total variance. The marked drop after the first component supports the retention of only one factor, reinforcing the unidimensionality of the analytic rubric ((Umlauft et al., 2019); (Watkins, 2018)). Together with the strong fit indices previously reported, the scree

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

plot provides additional visual confirmation that the data structure is best represented by a single latent factor.

Model Fit Analysis for Polytomous IRT Models Results

To identify the most appropriate polytomous item response theory (IRT) model for analyzing the constructed-response items scored using analytic coding, three commonly used models were compared: the Partial Credit Model (PCM), the Generalized Partial Credit Model (GPCM), and the Graded Response Model (GRM). These models were evaluated based on a set of statistical fit indices, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Sample-Size Adjusted BIC (SABIC), and Hannan–Quinn (HQ) (Wang et al., 2018). In addition, likelihood-based comparisons such as log-likelihood values and chi-square difference tests (ΔX^2) were used to assess whether the more complex models provided significantly better fit than the baseline PCM ((Wetzel & Carstensen, 2014); (Wind, 2022)). The summary of this model comparison is presented in Table 3.

Table 5. Comparison of model fit statistics for polytomous IRT models

Model Comparison	AIC	SABIC	HQ	BIC	Log-Likelihood	ΔX^2	df	*p*-value
PCM (Rasch)	20616.28	20680.95	20698.97	20830.18	-10261.14	—	—	—
GPCM	**20324.60**	**20418.16**	**20444.23**	**20634.07**	**-10094.30**	**333.68**	21	< .001
GRM (Graded)	20445.66	20539.22	20565.29	20755.13	-10154.83	-121.06	0	NaN

As shown in Table 5, the GPCM yielded the lowest values for AIC, SABIC, HQ, and BIC among the three models, indicating the best overall model fit to the analytic scoring data (Eckerly et al., 2022; Gos et al., 2020). The log-likelihood of the GPCM (-10094.30) is substantially higher than that of the PCM (-10261.14), and the chi-square difference test (ΔX^2 = 333.68, df = 21, $p < .001$) confirms that the GPCM provides a significantly better fit than the more restrictive PCM (de Ayala, 2009; Essen et al., 2017; Kim & Wilson, 2019). Meanwhile, the GRM did not offer a valid chi-square comparison in this context and exhibited poorer fit statistics overall.

To further examine model fit beyond information criteria and log-likelihood comparisons, a more robust fit evaluation was conducted using M2 statistics. The M2 analysis offers a limited-information goodness-of-fit approach suitable for multidimensional and polytomous models, particularly in the context of item response theory. It provides several fit indices including the M2 statistic with degrees of freedom and associated p-values, Root Mean Square Error of Approximation (RMSEA) with its 90% confidence interval, Standardized Root Mean Square Residual (SRMSR), Tucker-Lewis Index (TLI), and Comparative Fit Index (CFI). These indicators allow for a more nuanced evaluation of how well the models represent the response data across all items. The summary of model fit based on M2 analysis is presented in Table 6.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Table 6. Model fit statistics based on M2 analysis

Model Comparison	M2	df	*p*-value	RMSEA	RMSEA 90% CI	SRMSR	TLI	CFI
PCM (Rasch)	211.25	206	0.386	0.006	[0.000 – 0.017]	0.112	0.9997	0.9997
GPCM	170.85	185	0.764	0.000	[0.000 – 0.012]	**0.029**	**1.001**	**1.000**
GRM (Graded)	163.06	185	0.876	0.000	[0.000 – 0.009]	0.054	1.0012	1.000

As summarized in Table 6, all three models demonstrate acceptable fit based on the M2 analysis. However, the GPCM and Graded Response Model show slightly superior fit indices compared to the PCM. The GPCM, in particular, achieved the lowest SRMSR (0.029) and an RMSEA of 0.000 with a narrow confidence interval [0.000 – 0.012], indicating an excellent fit to the data (de Ayala, 2009; Essen et al., 2017; Felt et al., 2017). Additionally, both TLI and CFI values for the GPCM exceeded 1.000, reflecting a near-perfect model fit and reinforcing the flexibility and robustness of the model.

The PCM, while also performing well (RMSEA = 0.006, p = 0.386), showed a notably higher SRMSR (0.112), suggesting a less optimal representation of the data patterns (Bonifay & Cai, 2017; Lorenzo-Seva & Ferrando, 2023; Rhemtulla et al., 2020). These results further validate the selection of the GPCM for modeling the analytic rubric-based scoring of constructed-response items, as it captures item-level variability and response structure with greater fidelity ((Bonifay & Cai, 2017); (Eckerly et al., 2022); (Kim & Wilson, 2019); (Zhao & Hambleton, 2017)). This finding aligns with the theoretical expectation that analytic scoring approaches are best modeled using flexible IRT frameworks that accommodate varying discrimination parameters.

To evaluate the appropriateness of each item within the applied item response models, the S-X^2 item-level fit statistics were computed for all constructed-response items across the PCM, GPCM, and Graded Response models. The S-X^2 statistic provides a sensitive measure of item-level misfit by comparing observed and expected response patterns. A p-value below 0.05 indicates potential misfit between the item response data and the model expectations. Table 7 presents the S-X^2 values and corresponding p-values for each item under the three polytomous IRT models, along with a summary classification of item fit.

Table 7. Item fit summary (S-X^2 statistics)

No	Item	PCM - S-X^2(p-value)	Fit PCM	GPCM - S-X^2(p-value)	Fit GPCM	Graded - S-X^2(p-value)	Fit Graded
1	ITEM1I	50.50 (0.37)	Fit	49.34 (0.72)	Fit	44.12 (0.80)	Fit
2	ITEM1II	57.60 (0.38)	Fit	41.11 (0.60)	Fit	44.39 (0.46)	Fit
3	ITEM1III	62.93 (0.17)	Fit	51.71 (0.37)	Fit	49.69 (0.37)	Fit
4	ITEM1IV	73.25 (0.05)	Fit	59.69 (0.06)	Fit	57.49 (0.10)	Fit
5	ITEM1V	59.46 (0.25)	Fit	30.81 (0.82)	Fit	34.56 (0.71)	Fit
6	ITEM1VI	68.13 (0.05)	Fit	51.88 (0.17)	Fit	73.82 (0.01)	Misfit
7	ITEM2I	163.03 (0.00)	Misfit	75.82 (0.05)	Fit	77.41 (0.03)	Misfit
8	ITEM2II	56.55 (0.19)	Fit	51.11 (0.51)	Fit	52.12 (0.51)	Fit
9	ITEM2III	51.52 (0.49)	Fit	38.13 (0.82)	Fit	47.71 (0.53)	Fit
10	ITEM2IV	50.76 (0.52)	Fit	50.04 (0.51)	Fit	53.84 (0.37)	Fit
11	ITEM2V	60.42 (0.23)	Fit	39.45 (0.63)	Fit	45.14 (0.42)	Fit
12	ITEM2VI	53.36 (0.50)	Fit	47.80 (0.56)	Fit	48.60 (0.53)	Fit

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

No	Item	PCM - S-X²(p-value)	Fit PCM	GPCM - S-X²(p-value)	Fit GPCM	Graded - S-X²(p-value)	Fit Graded
13	ITEM2VII	56.36 (0.32)	Fit	55.54 (0.49)	Fit	54.57 (0.53)	Fit
14	ITEM2VIII	81.38 (0.01)	Misfit	52.53 (0.07)	Fit	48.36 (0.07)	Fit
15	ITEM2IX	52.01 (0.47)	Fit	50.13 (0.70)	Fit	51.51 (0.65)	Fit
16	ITEM2X	55.75 (0.37)	Fit	36.98 (0.69)	Fit	39.89 (0.56)	Fit
17	ITEM2XI	71.07 (0.03)	Misfit	61.62 (0.31)	Fit	66.02 (0.19)	Fit
18	ITEM2XII	53.34 (0.72)	Fit	58.14 (0.68)	Fit	65.08 (0.37)	Fit
19	ITEM3I	76.47 (0.01)	Misfit	65.60 (0.11)	Fit	73.32 (0.04)	Misfit
20	ITEM3II	57.35 (0.22)	Fit	46.72 (0.81)	Fit	48.78 (0.77)	Fit
21	ITEM3III	62.95 (0.14)	Fit	61.42 (0.29)	Fit	61.99 (0.27)	Fit
22	ITEM3IV	43.32 (0.77)	Fit	44.77 (0.75)	Fit	51.01 (0.51)	Fit

The results presented in Table 7 indicate that the vast majority of items demonstrate acceptable fit under all three models, particularly under the GPCM, which consistently yielded higher p-values and fewer misfitting items (Dimitrov & Luo, 2019; Zhao & Hambleton, 2017). Notably, ITEM2I and ITEM3I were flagged for misfit under the PCM model, while ITEM1VI, ITEM2I, and ITEM3I showed misfit under the Graded Response model. In contrast, the GPCM provided adequate fit for all but one item (ITEM2I at p = 0.05, borderline fit) ((Huen et al., 2023); (Wallmark et al., 2023)), supporting its flexibility in handling the variation in item discriminations inherent in analytic rubric scoring.

The reduced incidence of misfitting items under the GPCM reinforces the suitability of this model for assessing constructed-response data that are scored using detailed analytic criteria. These findings further validate the robustness of the GPCM framework for modeling complex, multidimensional performance tasks, such as those involved in the measurement of mathematical abstraction. This item-level diagnostic provides critical evidence for the appropriateness of model selection in subsequent validity analyses and scoring interpretation.

Results of Local Independence Assumption Testing for the GPCM Model

As part of the assumptions underlying item response theory, the principle of local independence posits that item responses should be conditionally independent given the latent trait. Violations of this assumption can bias parameter estimates and compromise the interpretability of the latent construct. To assess local independence among items, a residual-based analysis was conducted using both residual LD (Local Dependence) statistics and Q3 residual correlations. Table 6 summarizes the distribution of these residual indices to evaluate whether any substantial local dependence exists among item pairs.

Table 8. Summary of local independence diagnostics based on residual analysis

Statistic	Residual LD	Q3 Residuals
Minimum	-0.110	-0.207
1st Quartile	-0.044	-0.073
Median	0.037	-0.047
Mean	0.012	-0.043
3rd Quartile	0.054	-0.014
Maximum	0.104	0.084

As shown in Table 8, the distribution of residual LD values and Q3 residuals suggests that local dependence is minimal across the item set. The

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

maximum residual LD was 0.104, and the maximum Q3 residual was 0.084—both values well below commonly used thresholds (e.g., Q3 > 0.2) that would indicate problematic local dependence. Furthermore, the median and mean values for both indicators are near zero, reinforcing the conclusion that item responses are largely conditionally independent given the latent trait.

These findings support the validity of the local independence assumption and indicate that the scoring structure and item design do not lead to excessive redundancy or overlap in item content. Consequently, the reliability and interpretability of the model estimates and person scores are preserved, strengthening the credibility of the construct being measured.

Person Fit Analysis

In addition to evaluating item-level fit, it is crucial to examine person fit to ensure that respondents' answer patterns align with the expectations of the measurement model. Person fit analysis helps identify individuals whose response behaviors may deviate from model assumptions due to guessing, carelessness, disengagement, or other non-modelled factors. This study employed multiple indices for assessing person fit, including Outfit, Z-Outfit, Infit, Z-Infit, and the standardized Zh statistic. Table 9 presents the person fit statistics for the participants.

Table 9. Person fit analysis

Person	Outfit	Z-Outfit	Infit	Z-infit	Zh	Fit Decision
1	0.74153228	-0.05369	1.08552015	0.33361992	0.05191887	Fit
2	0.05901748	-0.824022	0.18706132	-1.0570694	1.03409003	Fit
3	1.02133911	0.20666634	1.209514	0.66139479	-0.2533299	Fit
4	1.08903817	0.50801491	1.09409062	0.52232603	-0.4484679	Fit
5	0.77995953	-0.1236926	1.21423064	0.59176505	0.00325445	Fit
6	0.38971037	-0.1004425	0.61970296	-0.2979855	0.4700821	Fit
7	0.46323566	-0.1058816	1.17757884	0.47365612	0.35681012	Fit
			(…some rows omitted for brevity)			
694	0.30774351	-0.2177306	0.51771705	-0.5587272	0.42055998	Fit
695	0.97660911	0.16515195	1.01364064	0.1731172	-0.2013791	Fit
696	1.00016916	0.27928069	1.28184005	0.6677195	-0.1277462	Fit
697	1.20780368	0.56395204	0.84071691	-0.3477306	-0.5231421	Fit
698	0.52275435	-0.0367469	0.79216079	-0.1032139	0.16917548	Fit
699	1.10863718	0.41539952	0.8611203	-0.4001978	-0.3194994	Fit
700	0.19841374	-0.4126616	0.5302778	-0.4314181	0.62787499	Fit

As shown in Table 9, all of respondents demonstrated adequate model fit across all person fit indices. Outfit and Infit values generally remained within the acceptable range (0.5 to 1.5), and Z-standardized statistics (Z-Outfit and Z-Infit) mostly hovered near zero, indicating no substantial misfit (Felt et al., 2017). Similarly, the Zh statistic did not flag any aberrant response patterns, with all values falling within conventional bounds ($|Zh| < 2.0$).

These results suggest that participants' responses are consistent with the underlying model assumptions and that the constructed-response items function well across the observed range of person abilities. The absence of person misfit enhances the credibility of the instrument by affirming that the data collected are psychometrically sound

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

and that person scores derived from the model are trustworthy for further interpretation.

Item Fit Statistics

To further evaluate the fit of the analytic-coded constructed-response items under the Generalized Partial Credit Model (GPCM), the study employed global item fit statistics using the S-X² chi-square statistic. This method examines the extent to which the observed response patterns for each item deviate from the expected response patterns generated by the model, while accounting for item degrees of freedom and sample size. Key indicators include the chi-square value (S-X²), degrees of freedom, RMSEA, and the associated p-value. These indicators collectively inform whether an item fits the GPCM assumptions at the global level. Table 10 presents the S-X² fit statistics for all 22 rubric-based item criteria analyzed.

Table 10. Global chi-square S-X² item fit statistics under the GPCM

Item	S-X²	*df*. S-X²	RMSEA. S-X²	*p*. S-X²	Fit Decision
ITEM1I	49.3383818	56	0	0.72328214	Fit
ITEM1II	41.107548	44	0	0.59631721	Fit
ITEM1III	51.714913	49	0.0089031	0.36826792	Fit
ITEM1IV	59.6882658	44	0.02258513	0.05750026	Fit
ITEM1V	30.8138138	39	0	0.82228169	Fit
ITEM1VI	51.8800172	43	0.01718835	0.16617191	Fit
ITEM2I	75.8236063	57	0.02173579	0.0484208	Fit
ITEM2II	51.1101562	52	0	0.50886948	Fit
ITEM2III	38.1332286	47	0	0.81842332	Fit
ITEM2IV	50.0359346	51	0	0.51192252	Fit
ITEM2V	39.4479742	43	0	0.62613661	Fit
ITEM2VI	47.7982036	50	0	0.56218868	Fit
ITEM2VII	55.5371688	56	0	0.49232636	Fit
ITEM2VIII	52.5257115	39	0.02227456	0.07255647	Fit
ITEM2IX	50.1328724	56	0	0.69547007	Fit
ITEM2X	36.9791666	42	0	0.69066602	Fit
ITEM2XI	61.6165287	57	0.0107642	0.31443851	Fit
ITEM2XII	58.1369216	64	0	0.68279384	Fit
ITEM3I	65.5974316	53	0.01844015	0.11471985	Fit
ITEM3II	46.7177478	56	0	0.8069997	Fit
ITEM3III	61.4194195	56	0.0117664	0.28808161	Fit
ITEM3IV	44.7658383	52	0	0.75143142	Fit

As shown in Table 10, all items met the global fit criteria under the GPCM framework. None of the items were flagged as misfitting, as indicated by their non-significant p-values ($p > 0.05$ for most items), low RMSEA values, and appropriate chi-square values relative to their degrees of freedom. Although a few items (e.g., ITEM2I, ITEM3I, ITEM3III) showed relatively lower p-values, they still fell within acceptable thresholds for model-based inferences and were retained due to theoretical alignment and overall model stability.

These results provide strong evidence for the global fit of the analytic-coded constructed-response items within the GPCM framework. The findings support the assumption that item responses reflect a consistent underlying latent trait—mathematical abstraction ability—and that the scoring rubric criteria are functioning in

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

accordance with the model's expectations. Consequently, the instrument demonstrates robustness and coherence at the item level, reinforcing its utility in measuring higher-order cognitive constructs using a polytomous IRT approach.

Item Parameter Estimates

Table 11 presents the item parameter estimates derived from the Generalized Partial Credit Model (GPCM) analysis applied to the constructed-response items. Each item is characterized by a discrimination parameter (a) and a set of step difficulty parameters (b1, b2, b3) depending on the number of score categories used in the analytic rubric. The column labeled *Location_gpcm* indicates the average item difficulty location, which is useful for interpreting the relative difficulty of each item on the same latent trait scale.

The discrimination parameter (a) reflects the item's sensitivity to differences in students' mathematical abstraction ability, while the step parameters (b1, b2, and b3) represent the thresholds between adjacent score categories. These values help evaluate how consistently and effectively the items differentiate among varying levels of the targeted construct.

Table 11. Parameter estimates of items based on the GPCM

Item	a	b1	b2	b3	Location_gpcm
ITEM1I	0.83649435	0.8011727	0.86498504		0.83307887
ITEM1II	1.9439452	1.04258889	1.25265561		1.14762225
ITEM1III	1.71123245	1.30256437	1.22438646		1.26347541
ITEM1IV	1.87676757	1.03094566	1.22308018		1.12701292
ITEM1V	2.21881735	1.2610972	1.16968743		1.21539231
ITEM1VI	1.98986167	1.27962448	1.48858594	1.32892511	1.36571184
ITEM2I	0.48379605	0.57996042	0.57425892		0.57710967
ITEM2II	1.04317914	0.81566444	1.10619112		0.96092778
ITEM2III	1.72287794	1.42863083	1.23373907		1.33118495
ITEM2IV	1.49687713	1.30970976	1.40352871		1.35661924
ITEM2V	1.97579522	1.28409012	1.20298946		1.24353979
ITEM2VI	1.53888261	1.24493384	1.30628178		1.27560781
ITEM2VII	1.07356418	1.01607822	1.26153799		1.13880811
ITEM2VIII	2.3408802	1.25619121	1.33558859		1.2958899
ITEM2IX	1.03894017	1.06146872	1.13430643		1.09788757
ITEM2X	1.91783591	1.4392154	1.21892791		1.32907166
ITEM2XI	1.02493436	1.21749925	0.94802527		1.08276226
ITEM2XII	1.14786205	1.06260339	1.2579922	1.36907662	1.22989074
ITEM3I	0.9602907	0.93936162	0.78746361		0.86341262
ITEM3II	1.03596842	0.94257129	1.17133895		1.05695512
ITEM3III	1.06724746	1.10431587	1.13940949		1.12186268
ITEM3IV	1.11649142	0.92455741	1.1251701		1.02486375

The parameter estimates shown in Table 11 reveal variation in both discrimination and difficulty levels across items and scoring criteria. Items such as ITEM1V and ITEM2VIII exhibit relatively high discrimination parameters (above 2.0), suggesting they are highly effective at distinguishing students with differing levels of mathematical abstraction. Conversely, items like ITEM2I and ITEM1I demonstrate lower discrimination, indicating more limited sensitivity to ability differences.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

The difficulty step parameters are generally ordered as expected, progressing from lower to higher values. However, a few irregularities in the step thresholds (e.g., ITEM2XI) may signal category functioning issues or inconsistent scoring. These anomalies warrant further examination, especially in relation to misfit indices and rubric validity.

The item locations, clustered mostly around the 0.58–1.37 logit range, suggest that the instrument is best targeted for students with moderate to slightly above-average levels of mathematical abstraction. This distribution is appropriate for diagnostic assessment and formative evaluation within the context of high school trigonometry instruction.

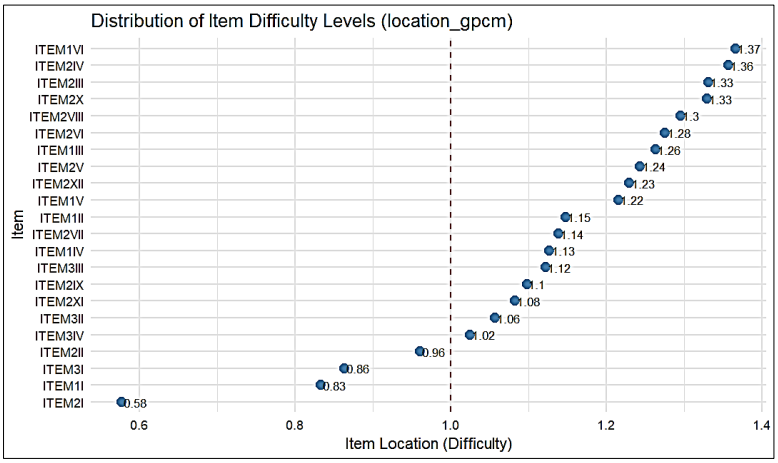


Figure 3. Item difficulty levels

Figure 3 illustrates the distribution of item difficulty levels based on the location_gpcm value from the GPCM model. The difficulty values range from approximately 0.58 to 1.37, with most items positioned above the reference line at 1.0, indicating a tendency toward moderate to high difficulty. This spread demonstrates sufficient variability in item difficulty, which is essential for effectively differentiating test-taker abilities. The distribution aligns with previous findings, supporting the unidimensionality of the analytic rubric and confirming that the items collectively measure a single latent construct of mathematical abstraction despite their varying difficulty levels.

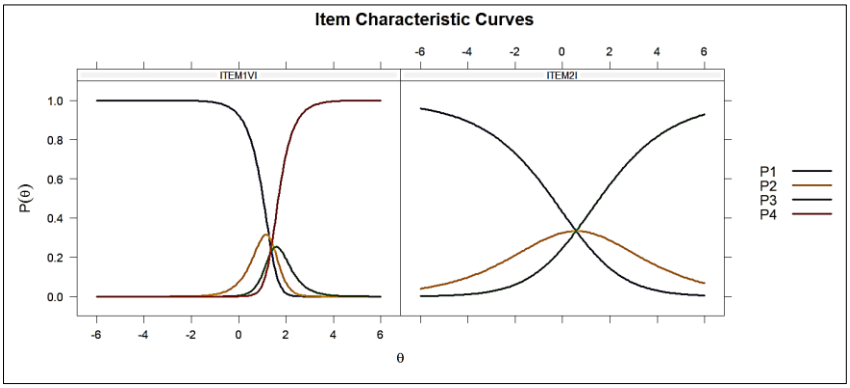


Figure 4. ICC of the most difficult (ITEM1VI) and easiest items (ITEM2I)

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Figure 4 displays the Item Characteristic Curves (ICCs) for the most difficult item (ITEM1VI) and the easiest item (ITEM2I) based on the GPCM model. The curves illustrate the probability of respondents selecting each score category across varying levels of the latent trait (θ). For ITEM1VI, the curves show that higher ability levels are required to achieve higher score categories, reflecting its greater difficulty. Conversely, ITEM2I's curves indicate that lower ability levels suffice to reach higher scores, consistent with its status as the easiest item. The clear separation and orderly progression of the category curves for both items confirm the appropriateness of the scoring rubric and support the unidimensionality of the construct being measured. These ICC patterns reinforce the validity of the analytic rubric in differentiating respondents' mathematical abstraction abilities across a range of item difficulties.

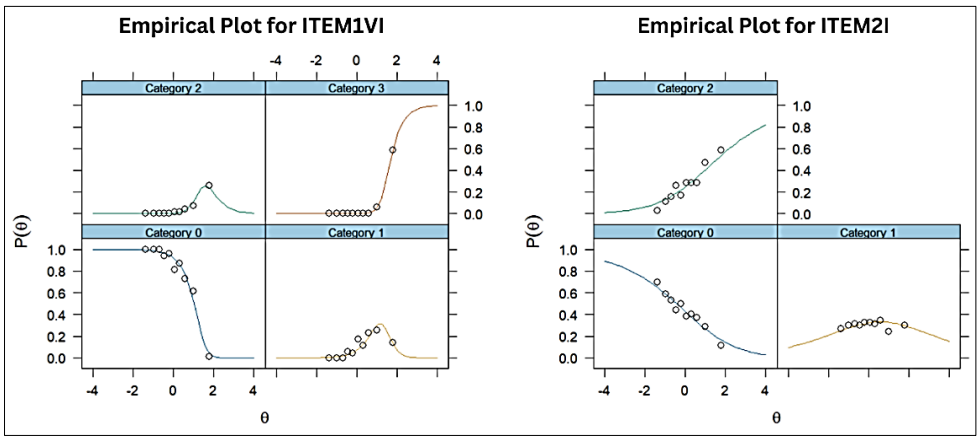


Figure 5. Empirical plot of the most difficult and easiest items

Figure 5 presents empirical plots illustrating the relationship between item scores and latent trait levels (θ) for the most difficult item (ITEM1VI) and the easiest item (ITEM2I). Each subplot displays the probability of respondents selecting specific score categories across the ability continuum. For ITEM1VI, the plots reveal that higher ability levels are necessary to achieve mid-to-high score categories, with clear transitions between categories indicating well-defined thresholds. In contrast, ITEM2I shows rapid increases in the probability of higher score categories at lower ability levels, reflecting its relative ease. The distinct separation and orderly progression of category probabilities in both items confirm the accuracy of the GPCM model in capturing item difficulty and category functioning. These empirical findings further validate the analytic rubric's capacity to differentiate respondents' abilities within a uni-dimensional measurement framework, reinforcing the robustness of the assessment instrument.

Respondent Ability

To examine how well the constructed-response items and their corresponding analytic scoring criteria reflect the latent trait of mathematical abstraction, person ability estimates were calculated using three different methods: Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP), and Expected A Posteriori (EAP). Each of these

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

estimation techniques offers distinct statistical advantages. MLE is commonly used but may be undefined for extreme response patterns, while MAP and EAP incorporate prior distributions, making them more robust in such cases. Table 10 presents the summary statistics of person ability estimates generated from each method.

Table 12. Summary of person ability estimates

Statistic	MLE	MAP	EAP
Minimum	-Inf	-1.364	-1.501
1st Quartile	-0.756	-0.586	-0.692
Median	0.005	0.005	-0.061
Mean	NaN	0.066	0.001
3rd Quartile	0.657	0.627	0.598
Maximum	Inf	2.918	3.039

As shown in Table 12, the MAP and EAP estimates provide meaningful person ability scores that are centered around zero, suggesting that the test items were appropriately targeted for the respondents. In contrast, the MLE estimates included undefined values (NaN and ±Inf), which commonly occur when participants answer all items correctly or incorrectly, highlighting a limitation of MLE in small samples or with constructed-response data. Overall, the MAP and EAP approaches offer more stable and interpretable estimates of mathematical abstraction ability in this analytic coding framework.

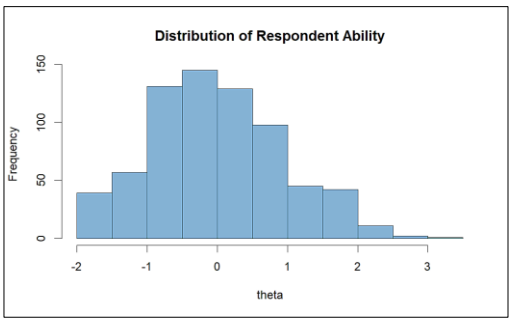


Figure 6. Distribution of respondent ability

Figure 6 illustrates the distribution of respondent abilities (θ) as a histogram, showing the frequency of individuals across different ability levels. The distribution appears approximately normal, with the majority of respondents clustered around the ability range of -1 to 1, peaking near zero. This indicates that most participants possess average to slightly below-average ability levels according to the measurement scale used. The spread of abilities extends from about -2 to just above 3, demonstrating a reasonable range of respondent proficiency within the sample. This distribution is important for understanding how well the test items target the population, as it suggests that the instrument is designed to assess a broad spectrum of abilities, with a concentration around the average level. These findings provide context for interpreting the test information function and item difficulty parameters presented in other figures, confirming that the test is well-aligned with the ability levels of the respondents.

Reliability

In evaluating the quality of the analytic rubric used for scoring mathematical abstraction, a Generalized Partial Credit Model (GPCM) analysis was conducted using the TAM package. One key focus of this analysis is the estimation of reliability, which reflects the consistency of the measurement instrument in differentiating student performance. The Expected A Posteriori (EAP) reliability index provides insight into how well the set of constructed-response items, scored using analytic criteria, functions as a coherent measurement tool. The summary of the GPCM estimation is presented in Table 13.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Table 13. Summary of GPCM Analysis

Component	Result
Number of Items	46 Categories (from 3 constructed-response items)
Total Parameters Estimated	46 thresholds (xsi parameters)
Range of Threshold Estimates	Min: 0.0470 (ITEM2I_Cat1)
	Max: 2.4181 (ITEM2X_Cat1)
EAP Reliability	**0.861**
Regression Coefficient	0.000
Person Variance	1.449

The EAP reliability value of 0.861 indicates that the analytic rubric yields highly consistent scores when used to measure students' mathematical abstracttion. This level of reliability suggests that the rubric is sufficiently precise in capturing true differences in students' abilities, minimizing measurement error. High reliability is essential in supporting valid interpretations of students' performance, especially in constructed-response tasks where scoring can be more subjective. These findings confirm the robustness of the analytic rubric as a dependable measurement instrument within the GPCM framework.

Item Information and Standard Errors

Factor loadings (F1), communalities (h^2), and standard errors (SE) were examined to assess how well each item criterion contributes to the latent trait of mathematical abstraction. Table 14 presents the statistical outputs for all 22 scored criteria, labeled by item number and analytic coding criterion (e.g., ITEM1I denotes Item 1, Criterion I).

Table 14. Item information, communality, and standard errors from GPCM model

Item	Factor Loading (F1)	Communality (h^2)	SE of F1
ITEM1I	0.672	0.451	0.037
ITEM1II	0.903	0.816	0.016
ITEM1III	0.880	0.775	0.020
ITEM1IV	0.897	0.805	0.017
ITEM1V	0.923	0.853	0.014
ITEM1VI	0.935	0.875	0.012
ITEM2I	0.464	0.216	0.047
ITEM2II	0.749	0.561	0.031
ITEM2III	0.882	0.777	0.020
ITEM2IV	0.851	0.725	0.023
ITEM2V	0.906	0.821	0.016
ITEM2VI	0.858	0.736	0.022
ITEM2VII	0.758	0.575	0.031
ITEM2VIII	0.930	0.866	0.013
ITEM2IX	0.748	0.559	0.032
ITEM2X	0.901	0.812	0.018
ITEM2XI	0.743	0.552	0.032
ITEM2XII	0.836	0.699	0.023
ITEM3I	0.721	0.520	0.033
ITEM3II	0.747	0.558	0.032
ITEM3III	0.757	0.572	0.031
ITEM3IV	0.771	0.594	0.030

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

As shown in Table 14, most analytic coding criteria demonstrated strong factor loadings, particularly ITEM1VI (0.935), ITEM2VIII (0.930), and ITEM1V (0.923), indicating high alignment with the latent trait being measured. The associated communalities (h^2) also reflect substantial shared variance with the underlying factor, further supporting the construct validity of the analytic rubric. Standard errors (SE) for most items were low, especially for the highly loading items, suggesting precise parameter estimates. Collectively, these findings suggest that the analytic coding rubric applied to the constructed-response items yields reliable and valid indicators of students' mathematical abstraction abilities, justifying its use in performance-based assessments aligned with the GPCM framework.

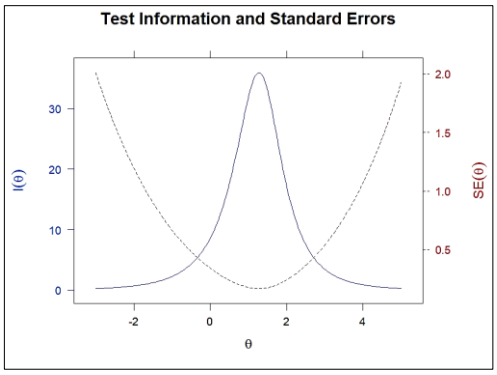


Figure 7. Test information and standard errors

Figure 7 depicts the test information function (solid blue line) and the standard error of measurement (dashed red line) plotted against respondent ability levels (θ). The test information curve reaches its maximum near an ability level of approximately 1.5, which indicates that the test items tend to have a higher difficulty level. This makes the instrument particularly suitable for measuring respondents with slightly above-average ability. Correspondingly, the standard error is minimized around this point, reflecting the highest measurement precision. As ability levels move away from this peak in either direction, the test information decreases and the standard error increases, indicating less precise measurement for respondents with very low or very high abilities. This pattern suggests that the instrument, modeled under the GPCM framework, is optimally reliable for assessing individuals with moderate to moderately high ability levels. These findings align well with the distribution of respondent abilities shown in Figure 6, confirming that the test is well-targeted to the ability range of the sample population.

The results of this study affirm the appropriateness of the Generalized Partial Credit Model (GPCM) in analyzing constructed-response assessments (CRA) that are scored using analytic rating rubrics—specifically designed to capture varying levels of mathematical abstraction in trigonometric ratio tasks. Unlike holistic scoring, analytic rubrics allow for the decomposition of student responses into distinct cognitive criteria, each scored independently. In this study, each rubric indicator was designed with different scoring levels (ranging from 0–3 or 0–4 points), tailored to the cognitive complexity of the expected student reasoning. As highlighted by Hamhuis et al. (2020), GPCM is particularly effective for assessments involving polytomous data—such as analytic rubrics—where student responses demonstrate varied levels of completeness and cognitive quality. In this study, GPCM provided strong item and person fit, confirming that the analytic rubric is capable of capturing the nuanced gradations of abstraction skill across student responses.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

The discrimination and threshold parameters indicate that the items effectively differentiate between students at various levels of abstraction ability, aligning with Bürkner et al. (2018) emphasis on designing assessment items that target higher-order thinking skills (HOTS). This is especially critical for trigonometric ratio problems, which demand conceptual reasoning, representational fluency, and the ability to abstract functional relationships. The findings suggest that each rubric-coded criterion successfully represents a continuum of mathematical abstraction, providing strong construct validity.

From a psychometric perspective, the high reliability indices and satisfactory separation values provide further evidence of the rubric's measurement precision. Wind's (2022) work supports this interpretation, noting that GPCM—especially when used in conjunction with MMLE estimation—can reliably evaluate rating scales, such as those used in CRA. The stability of person estimates generated via EAP and MAP methods in this study confirms the rubric's robustness in producing replicable and meaningful scores.

Moreover, the test information curve shows that the assessment provides optimal precision for students with moderate to high levels of abstraction ability. This insight is pedagogically valuable: it reveals the rubric's diagnostic potential for identifying students who are ready for more complex problem-solving versus those who may need scaffolded support. As Mustangin and Setiawan (2021) argue, conceptual mastery in trigonometry is a strong predictor of success in mathematical tasks; the current findings validate that analytic rubrics, modeled with GPCM, can localize specific breakdowns in student reasoning and inform instructional adjustments.

Additionally, consistent with findings by Setiawan (2022) and Usman and Hussaini (2017), the model captures common student difficulties in conceptualizing trigonometric ratios—especially when reasoning about angle magnitudes, ratios, and graphical representations. GPCM offers a framework for systematically diagnosing such misconceptions by evaluating the distribution and progression of rubric-based responses. This adds a diagnostic layer that surpasses traditional multiple-choice assessments, which often obscure students' thought processes.

The study also reinforces Zhou and Huggins-Manley's (2020) argument about the importance of addressing item-level missing data in large-scale assessments. Although missingness was minimal in this study, GPCM's capacity to handle incomplete responses ensures that future applications in broader settings remain methodologically sound.

Furthermore, the affective-motivational dimension cannot be overlooked. Gilbert (2016) assertion that motivation shapes performance in CRA is particularly relevant in trigonometry, where task complexity can either engage or discourage students. The variability in response quality observed in this study may reflect differing levels of self-efficacy or persistence. Integrating motivation-related indicators into future GPCM applications may yield even richer insights.

Lastly, the analytical precision of GPCM allows for domain-specific refinement of assessment tools. As Buchholz and Hartig (2017) suggest, GPCM enables researchers to evaluate invariance across different populations

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

and contexts. Applied here, it opens opportunities to examine whether abstraction performance varies across instructional models, school types, or curriculum tracks—thus bridging psychometric rigor with instructional relevance.

CONCLUSIONS

This study provides robust empirical evidence supporting the validity and reliability of an analytic rubric for assessing mathematical abstraction through constructed-response items within the Generalized Partial Credit Model (GPCM) framework. The item-level analyses demonstrated strong discrimination, appropriate step difficulty ordering, and clear item characteristic curve (ICC) patterns, confirming that the rubric effectively differentiates students across varying levels of abstraction ability. Global and local item fit statistics further validated the model assumptions of unidimensionality and local independence, while person fit indices indicated consistent and credible response behavior.

The ability estimates generated through MAP and EAP methods proved stable and interpretable, with reliability coefficients exceeding conventional thresholds. These findings highlight the analytic rubric's capability to provide precise measurement with minimized error, especially in the context of performance-based assessments. The distribution of person abilities and test information curves revealed that the instrument is well-targeted for students with moderate to slightly above-average abstraction abilities—an important consideration for instructional alignment and formative evaluation.

Additionally, the high factor loadings and communalities among rubric-coded criteria reinforce the construct validity of the abstraction construct as measured. Together, these findings confirm that the analytic rubric, when scored using well-defined criteria and modeled through GPCM, serves as a psychometrically sound tool for capturing students' mathematical abstraction in the domain of trigonometric ratio. This contributes to the advancement of measurement practices for higher-order thinking in mathematics education and supports the shift toward more authentic, rubric-based assessment frameworks.

REFERENCES

Amelia, R., Listiaji, P., Dewi, N. R., Heriyanti, A. P., Atmaja, B. D., Shoba, T. M., & Sajidi, I. (2024). Developing and Validating a Rubric for Measuring Skills in Designing Science Experiments for Prospective Science Teachers. *Jurnal Inovasi Pendidikan Ipa*, *10*(1), 32–46. https://doi.org/10.21831/jipi.v10i1.65853

Angraini, L. M. (2018). Pengaruh Concept Attainment Model Terhadap Disposisi Berpikir Kritis Matematis Mahasiswa. *JNPM (Jurnal Nasional Pendidikan Matematika)*, *2*(2), 284. https://doi.org/10.33603/jnpm.v2i2.1473

Attali, Y., Laitusis, C., & Stone, E. (2016). Differences in Reaction to Immediate Feedback and Opportunity to Revise Answers for Multiple-Choice and Open-Ended Questions. *Educational and Psychological Measurement*, *76*(5), 787–802. https://doi.org/10.1177/0013164415612548

Bonifay, W., & Cai, L. (2017). On the Complexity of Item Response

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Theory Models. *Multivariate Behavioral Research*, *52*(4), 465–484. https://doi.org/10.1080/00273171.2017.1309262

Buchholz, J., & Hartig, J. (2017). Comparing Attitudes Across Groups: An IRT-Based Item-Fit Statistic for the Analysis of Measurement Invariance. *Applied Psychological Measurement*, *43*(3), 241–250. https://doi.org/10.1177/0146621617748323

Bürkner, P., Schwabe, R., & Holling, H. (2018). Optimal Designs for the Generalized Partial Credit Model. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 271–293. https://doi.org/10.1111/bmsp.12148

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory* (D. A. Kenny, Ed.). The Guilford Press.

Dimitrov, D. M., & Luo, Y. (2019). A Note on the D-Scoring Method Adapted for Polytomous Test Items. *Educational and Psychological Measurement*, *79*(3), 545–557. https://doi.org/10.1177/0013164418786014

Eckerly, C., Jia, Y., & Jewsbury, P. (2022). Technology-Enhanced Items and Model–Data Misfit. *ETS Research Report Series*, *2022*(1), 1–16. https://doi.org/10.1002/ets2.12353

Edimuslim, E. (2022). Analisis Kemampuan Abstraksi Matematis Siswa Sekolah Menengah Pertama Ditinjau Dari Gaya Belajar Tipe Kolb. *Suska Journal of Mathematics Education*, *8*(1), 39. https://doi.org/10.24014/sjme.v8i1.16831

Erni, E., Ma'rufi, M., & Ilyas, M. (2022). Pengaruh Kemadirian Belajar Terhadap Kemampuan Berpikir Kreatif Matematika Siswa. *Kognitif Jurnal Riset Hots Pendidikan Matematika*, *2*(1), 53–61. https://doi.org/10.51574/kognitif.v2i1.386

Essen, C. B., Idaka, I. E., & Metibemu, M. A. (2017). Item Level Diagnostics and Model - Data Fit in Item Response Theory (IRT) Using BILOG - MG v3.0 and IRTPRO v3.0 Programmes. *Global Journal of Educational Research*, *16*(2), 87. https://doi.org/10.4314/gjedr.v16i2.2

Faisal, A. F., Lambertus, L., & Baharuddin, B. (2020). Pengaruh Kemandirian Belajar Matematik Siswa Terhadap Kemampuan Berpikir Kreatif Matematis Siswa SMA Negeri 03 Bombana. *Jurnal Pembelajaran Berpikir Matematika (Journal of Mathematics Thinking Learning)*, *5*(2). https://doi.org/10.33772/jpbm.v5i2.15749

Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using Person Fit Statistics to Detect Outliers in Survey Research. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00863

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *49*(2), 78–88. https://doi.org/10.1037/cbs0000069

Gibson Jr., T. O., Morrow, J. A., & Rocconi, L. M. (2020). A Modernized Heuristic Approach to Robust Exploratory Factor Analysis. *The Quantitative Methods for Psychology*, *16*(4), 295–307. https://doi.org/10.20982/tqmp.16.4.p295

Gilbert, M. C. (2016). Relating aspects of motivation to facets of mathematical

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

competence varying in cognitive demand. *The Journal of Educational Research*, *109*(6), 647–657. https://doi.org/10.1080/00220671.2015.1020912

Gos, E., Sagan, A., Skarżyński, P. H., & Skarżyński, H. (2020). Improved Measurement of Tinnitus Severity: Study of the Dimensionality and Reliability of the Tinnitus Handicap Inventory. *Plos One*, *15*(8), e0237778. https://doi.org/10.1371/journal.pone.0237778

Hamhuis, E. R., Glas, C. A. W., & Meelissen, M. R. (2020). Tablet Assessment in Primary Education: Are There Performance Differences Between TIMSS' Paper-and-pencil Test and Tablet Test Among Dutch Grade-four Students? *British Journal of Educational Technology*, *51*(6), 2340–2358. https://doi.org/10.1111/bjet.12914

Hawai, M. F. (2021). Proses Berpikir Matematis Siswa Dalam Menyelesaikan Soal PISA Kategori HOTS Dan Scaffoldingnya. *Mathedunesa*, *10*(1), 95–109. https://doi.org/10.26740/mathedunesa.v10n1.p95-109

Huen, J. M. Y., Yip, P. S. F., Osman, A., & Leung, A. N. M. (2023). Item Response Theory and Differential Item Functioning Analyses With the Suicidal Behaviors Questionnaire–Revised in US and Chinese Samples. *Crisis*, *44*(2), 108–114. https://doi.org/10.1027/0227-5910/a000837

Karakuş, G., & Ocak, G. (2022). The Implementation of Cooperative Problem-Solving Rubric Towards Turkish Fourth Grade Students. *Mimbar Sekolah Dasar*, *9*(1), 1–23. https://doi.org/10.53400/mimbar-sd.v9i1.39390

Khairunnisa, I., Ariyanto, L., & Endahwuri, D. (2021). Analisis Berpikir Kreatif Matematis Ditinjau Dari Motivasi Belajar Siswa. *Imajiner Jurnal Matematika Dan Pendidikan Matematika*, *3*(6), 527–534. https://doi.org/10.26877/imajiner.v3i6.8087

Kim, J., & Wilson, M. (2019). Polytomous Item Explanatory Item Response Theory Models. *Educational and Psychological Measurement*, *80*(4), 726–755. https://doi.org/10.1177/0013164419892667

Kuo, B.-C., Chen, C.-H., Yang, C.-W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology*, *36*(6), 1115–1133. https://doi.org/10.1080/01443410.2016.1166176

Lorenzo-Seva, U., & Ferrando, P. J. (2023). A Simulation-Based Scaled Test Statistic for Assessing Model-Data Fit in Least-Squares Unrestricted Factor-Analysis Solutions. *Methodology*, *19*(2), 96–115. https://doi.org/10.5964/meth.9839

Mustangin, M., & Setiawan, Y. E. (2021). Pemahaman Konsep Mahasiswa Semester Satu Pada Mata Kuliah Trigonometri. *Jurnal Elemen*, *7*(1), 98–116. https://doi.org/10.29408/jel.v7i1.2773

Navas-López, E. A. (2024). Confirmatory Factor Analysis of a Rubric for Assessing Algorithmic Thinking on Undergraduate Students. *Cuadernos De Investigación Educativa*, *15*(2). https://doi.org/10.18861/cied.2024.15.2.3797

Ngu, B. H., & Phan, H. P. (2020). Learning to Solve Trigonometry Problems That Involve Algebraic Transformation Skills via Learning by Analogy and Learning by

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

Comparison. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.558773

Ocy, D. R., Rahayu, W., & Makmuri, M. (2023). Rasch Model Analysis: Development Of Hots-Based Mathematical Abstraction Ability Instrument According To Riau Islands Culture. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *12*(4), 3542–3560.

Prihono, E. W., Lapele, F., Jumaeda, S., Sukadari, S., & Nurjanah, S. (2022). *EFA of Pedagogic Competence Instrument to Measure Teacher Performance*. https://doi.org/10.2991/assehr.k.220129.059

Rhemtulla, M., Bork, R. v., & Borsboom, D. (2020). Worse Than Measurement Error: Consequences of Inappropriate Latent Variable Measurement Models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2020). *An R Toolbox for Score-Based Measurement Invariance Tests in IRT Models*. https://doi.org/10.31234/osf.io/r9w34

Setiawan, Y. E. (2022). Prospective Teachers Representations in Problem Solving of Special Angle Trigonometry Functions Based on the Level of Ability. *Infinity Journal*, *11*(1), 55. https://doi.org/10.22460/infinity.v11i1.p55-76

Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics*, *9*(1), 4–11. https://doi.org/10.12691/ajams-9-1-2

Sukmawan, I., Sridana, N., & Novitasari, D. (2022). Hubungan Konsep Diri Terhadap Kemampuan Berpikir Logis Matematis Siswa SMP Negeri 18 Mataram Tahun Pelajaran 2021/2022. *Jurnal Ilmiah Profesi Pendidikan*, *7*(3b), 1564–1571. https://doi.org/10.29303/jipp.v7i3b.816

Syarifudin, M. T., Ratnaningsih, N., & Ni'mah, K. (2021). Analisis Kemampuan Abstraksi Matematis dalam Pembelajaran Matematika di MAN 1 Tasikmalaya. *MUST: Journal of Mathematics Education, Science and Technology*, *6*(2), 231. https://doi.org/10.30651/must.v6i2.7461

Umlauft, M., Placzek, M., Konietschke, F., & Pauly, M. (2019). Wild Bootstrapping Rank-Based Procedures: Multiple Testing in Nonparametric Factorial Repeated Measures Designs. *Journal of Multivariate Analysis*, *171*, 176–192. https://doi.org/10.1016/j.jmva.2018.12.005

Usman, M. H., & Hussaini, M. M. (2017). Analysis of Students' Error in Learning of Trigonometry Among Senior Secondary School Students in Zaria Metropolis, Nigeria. *Iosr Journal of Mathematics*, *13*(02), 01–04. https://doi.org/10.9790/5728-1302040104

Wallmark, J., Ramsay, J. O., Li, J., & Wiberg, M. (2023). Analyzing Polytomous Test Data: A Comparison Between an Information-Based IRT Model and the Generalized Partial Credit Model. *Journal of Educational and Behavioral Statistics*, *49*(5), 753–779. https://doi.org/10.3102/10769986231207879

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting Aberrant Behavior and Item Preknowledge: A Comparison of Mixture Modeling Method and Residual Method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i4.12976

https://doi.org/10.3102/1076998618767123

Watkins, M. W. (2018). Exploratory Factor Analysis: A Guide to Best Practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

Watson, J. C. (2017). Establishing Evidence for Internal Structure Using Exploratory Factor Analysis. *Measurement and Evaluation in Counseling and Development*, *50*(4), 232–238. https://doi.org/10.1080/07481756.2017.1336931

Wetzel, E., & Carstensen, C. H. (2014). Reversed Thresholds in Partial Credit Models. *Assessment*, *21*(6), 765–774. https://doi.org/10.1177/1073191114530775

Wind, S. A. (2022). Detecting Rating Scale Malfunctioning With the Partial Credit Model and Generalized Partial Credit Model. *Educational and Psychological Measurement*, *83*(5), 953–983. https://doi.org/10.1177/0013164422116292

Yanti, N. F., & Wijaya, A. (2023). Meta-Analisis: Pengaruh Penerapan Model Pembelajaran Problem-Based Learning Terhadap Kemampuan Berpikir Kritis Matematis Siswa. *Aksioma Jurnal Program Studi Pendidikan Matematika*, *12*(1), 1213. https://doi.org/10.24127/ajpm.v12i1.6750

Yao, L., & Schwarz, R. D. (2006). A Multidimensional Partial Credit Model With Associated Item and Test Statistics: An Application to Mixed-Format Tests. *Applied Psychological Measurement*, *30*(6), 469–492. https://doi.org/10.1177/0146621605284537

Zhao, Y., & Hambleton, R. K. (2017). Practical Consequences of Item Response Theory Model Misfit in the Context of Test Equating With Mixed-Format Test Data. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00484

Zhou, S., & Huggins-Manley, A. C. (2020). The Performance of the Semigeneralized Partial Credit Model for Handling Item-Level Missingness. *Educational and Psychological Measurement*, *80*(6), 1196–1215. https://doi.org/10.1177/0013164420918392