Item Analysis of Mathematical Critical Thinking Skills Using the Rasch Model in Elementary School Students

Yogina Pratiwi*, Prihantini, Abubakar

Universitas Pendidikan Indonesia, Bandung, Indonesia
*Corresponding author, email: yoginapratiwi@upi.edu

https://doi.org/10.17977/um065.v7.i3.2027.4

Article history
Submitted: 11 July 2026
Revised: 25 August 2026
Accepted: 28 August 2026
Published: 1 September 2026

Keywords
Critical thinking
Elementary school students
Facione
Mathematics
Rasch model

Abstract
The validity and reliability of educational assessment instruments are fundamental prerequisites for accurate measurement of student competencies. This study evaluates the psychometric quality of a 12-item mathematical critical thinking test grounded in the Facione framework, encompassing Interpretation, Analysis, Inference, and Evaluation, administered to 49 Grade V elementary school students (n = 49) using the Rasch measurement model via Winsteps software. Four psychometric dimensions were assessed simultaneously: (1) construct validity through item fit statistics (Outfit MNSQ, Outfit ZSTD, Pt Measure Corr); (2) internal consistency and reliability (KR-20, Person and Item Reliability, Separation Indices); (3) item discrimination (Pt Measure Corr classification); and (4) item difficulty hierarchy (logit scale measures and Wright Person–Item Map). Results show that 10 of 12 items (83.3%) met Rasch fit criteria, yielding construct-valid measures for all four critical thinking indicators. KR-20 = 0.79 (Good), Person Reliability = 0.79–0.82 (Adequate–Good), and Item Reliability = 0.95 (Excellent). Person Separation (1.96–2.12) confirmed differentiation of ≥2 ability strata; Item Separation (4.37–4.61) confirmed a well-differentiated difficulty hierarchy. All retained items exhibited Moderate difficulty (logit range: −0.66 to +0.23), confirming optimal targeting of the respondent population. Two items failed due to low Point Measure Correlation (0.21–0.22) and were excluded. The final 10-item instrument meets international psychometric standards and is recommended for use in authentic mathematical critical thinking assessment.

1. Introduction

Mathematics achievement in the early years of schooling is a robust predictor of students' longer-term academic trajectories, with early numeracy skills and confidence in mathematics contributing directly to fourth-grade mathematics outcomes in large-scale international assessment data (Chang, 2023). Within this broader foundation, critical thinking is widely recognised as the most indispensable cognitive skill in the 21st-century knowledge economy. Facione (2015) defines critical thinking as purposive, self-regulating judgement that produces interpretation, analysis, evaluation, inference, explanation, and self-regulation based on evidential, conceptual, methodological, and contextual considerations. A three-year longitudinal study of upper elementary students found that critical thinking and academic achievement predict one another bidirectionally, independent of general cognitive ability, underscoring that critical thinking is not merely a by-product of learning but an active driver of it from an early age (Lv, Zhou, & Ren, 2025). In mathematics education, critical thinking is not merely a desirable addition but an epistemically essential outcome: mathematics offers an unparalleled context for developing the Interpretation of data and problem representations, Analysis of logical relationships, Evaluation of solution validity, and Inference from mathematical evidence (Dwyer, Hogan, & Stewart, 2014). The centrality of critical thinking in mathematics is affirmed globally, including in Indonesia's *Kurikulum Merdeka*, which explicitly positions critical thinking as a priority competency for elementary school students (Kementerian Pendidikan, 2022).

Despite this recognition, empirical evidence reveals a persistent gap in mathematical critical thinking performance among Indonesian elementary school students. PISA 2022 reported Indonesian students' mean mathematics score at 366 points, 106 points below the OECD average of 472, with only 18% of students reaching at least Level 2 proficiency compared with 69% across OECD countries, indicating that the large majority performed below the threshold associated with basic mathematical reasoning rather than merely at Levels 1–2 (OECD, 2023). These findings indicate not only a teaching gap but also a serious measurement gap: psychometrically inadequate critical thinking assessment instruments will generate data that misrepresent students' actual critical thinking capacity, rendering designed learning interventions ineffective.

ISSN: 2797-3174

This measurement gap is rooted in the dominance of Classical Test Theory (CTT) in Indonesian educational research instrument development. CTT produces item statistics, difficulty index (p-value), point-biserial correlation, Cronbach's alpha, that are fundamentally sample- and test-dependent (Hambleton & Jones, 1993; Guler & Uyanik, 2021). This limitation is particularly critical for critical thinking instruments, where item complexity and construct breadth amplify the sensitivity of statistics to sample characteristics (Bond, Yan, & Heene, 2021). In response, the Rasch measurement model offers a methodologically superior solution: by placing persons and items on a shared probabilistic logit scale, it generates sample-invariant item parameters and test-invariant person ability estimates (Bond, Yan, & Heene, 2021). For dichotomously scored items, the probability of a correct response is expressed as $P(X=1|\theta, \delta) = \exp(\theta-\delta) / [1 + \exp(\theta-\delta)]$, where θ is person ability and δ is item difficulty on the same logit scale. The adoption of Rasch analysis for critical thinking instrument validation continues to grow internationally, spanning domains from reading comprehension (Santos et al., 2016) to mathematics construct modelling (Rittle-Johnson, Matthews, Taylor, & McEldoon, 2011), supported by widely used methodological guides for conducting and interpreting Rasch-based psychometric evaluation (Tavakol & Dennick, 2013).

A systematic review of instructional interventions for cultivating critical thinking through mathematics found that studies consistently struggle to disentangle the specific cognitive dimensions of critical thinking being measured, and called for assessment instruments with clearer, more explicit operational grounding in an established critical thinking framework (Wang & Abdullah, 2024). Several researchers have conducted psychometric validation of critical thinking or higher-order thinking (HOTS) instruments in mathematics and science education using Rasch modelling. Sujatmika et al. (2025) applied a polytomous Rasch model to validate a six-dimension critical thinking instrument for junior high school science students and found that Inference and Evaluation items were consistently more difficult than Analysis and Interpretation items, a hierarchy broadly consistent with the Facione (2015) framework, although the study covered a science rather than a mathematics context. At the elementary level specifically, Mapeala and Siew (2015) developed and validated a science critical thinking test for fifth-grade students, while Rittle-Johnson et al. (2011) used a construct-modelling approach grounded in item response theory to assess elementary and middle-school students' knowledge of mathematical equivalence, illustrating how a Wright-map-based approach can diagnose specific conceptual gaps in mathematics. However, psychometric validation using Rasch analysis, four-dimensional coverage explicitly grounded in Facione's framework, and explicit item-difficulty mapping via a Wright Person–Item Map remain uncommon in existing mathematics-specific critical thinking instruments. Beyond these studies, the broader body of Rasch-based validation work in mathematics education has tended to report only partial psychometric indicators, limited explicit grounding in an operational critical thinking framework, and inconsistent inclusion of the Wright Person–Item Map.

A review of these studies reveals three consistent and unmet methodological gaps in the Rasch-based critical thinking mathematics instrument validation literature in Indonesia. First, most studies report only partial psychometric indicators typically MNSQ and KR-20, without integrating discrimination (Pt Measure Corr) and item difficulty mapping (logit measures) into a unified four-dimensional analysis. Second, many studies lack an explicit grounding in an operational critical thinking theoretical framework, particularly Facione's (2015) taxonomy of Interpretation, Analysis, Inference, and Evaluation, so the construct validity claimed cannot be traced to a valid critical thinking conceptual framework. Third, the Wright Person–Item Map, the most diagnostically informative output of Rasch analysis for visualising the alignment between student ability and item difficulty, is rarely included in Indonesian critical thinking instrument studies.

These three gaps form the basis and direction of this study. This study conducts a comprehensive Rasch-based psychometric validation of a 12-item mathematical critical thinking test instrument developed based on the Facione (2015) framework, administered to 49 elementary school students using Winsteps version 3.73. Validation was conducted simultaneously across four psychometric dimensions: (1) construct validity through item fit statistics (Outfit MNSQ, Outfit ZSTD, and Pt Measure Corr); (2) internal consistency and reliability through KR-20, Person and Item Reliability, and Separation Indices; (3) item discrimination through Pt Measure Corr classification; and (4) item difficulty hierarchy through logit measures and the Wright Person–Item Map. With an integrated four-dimensional approach explicitly grounded in the Facione (2015) framework, this study seeks to address the gaps identified in previous studies.

Based on the background and research gaps, the objectives of this study are: (1) to identify items meeting the Rasch model fit criteria from the 12-item mathematical critical thinking instrument based on the Facione (2015) framework; (2) to evaluate the level of internal consistency, person reliability, and item reliability of the instrument; (3) to analyse the discriminating power of each item based on Pt Measure Corr classification; and (4) to map the distribution of item difficulty along the logit hierarchy through the Wright Person–Item Map. Operationally, this study addresses four research questions: (RQ1) Which items meet the Rasch model fit criteria? (RQ2) What is the level of internal consistency, person reliability, and item reliability? (RQ3) What is the discriminating power of each item? (RQ4) How are items distributed across the difficulty hierarchy?

ISSN: 2797-3174

2. Method

2.1. Research Approach and Design

This study employed a quantitative approach with a psychometric research design. The quantitative approach was selected because the research objective is to produce objective, replicable numerical measures of the psychometric quality of the instrument, encompassing validity, reliability, discrimination, and item difficulty levels. Within this framework, the psychometric design specifically aims to evaluate whether an instrument accurately, consistently, and without sample bias measures the intended construct (Bond, Yan, & Heene, 2021). The Rasch measurement model was selected as the analytical framework due to its superiority over Classical Test Theory in generating sample-invariant item parameters and test-invariant respondent ability estimates (Boone & Noltemeyer, 2017). Winsteps version 3.73 was used as the standard implementation tool for Rasch analysis in this study.

2.2. Research Setting and Timeframe

The study was conducted at a public elementary school (referred to hereafter as public elementary school X) in a regency in Southeast Sulawesi Province, Indonesia; the school's identity is withheld to protect participant confidentiality. This location was selected on the basis that public elementary school X has implemented the *Kurikulum Merdeka*, which explicitly positions critical thinking as a priority competency, making it a relevant context for validating a mathematical critical thinking instrument. Data collection was carried out in the even semester of the 2025/2026 academic year under standardised learning conditions.

2.3. Research Subjects

The research subjects were 49 Grade V students of public elementary school X, selected using purposive sampling. This technique involves determining samples based on specific considerations and criteria established by the researcher in accordance with the study objectives. Purposive sampling was chosen because the study aims to validate a mathematical critical thinking instrument with students who have relevant prior learning experience, requiring a sample with specific characteristics that cannot be obtained through random sampling techniques. Sampling criteria included: (1) students had completed the mathematics content relevant to the critical thinking constructs being measured; (2) students were present at the time of data collection; and (3) answer sheets were completed in full. Ethical considerations were addressed, including obtaining permission from the school principal, parental or guardian consent, and anonymisation of respondent data in accordance with applicable educational research guidelines.

2.4. Research Instrument

The research instrument consisted of 12 mathematical word problems referencing addition and subtraction of whole numbers, scored dichotomously (1 = correct, 0 = incorrect). The items were developed to measure four critical thinking skill indicators based on the Facione (2015) framework: Interpretation, Analysis, Inference, and Evaluation. Each indicator was represented by three items in a balanced manner: Interpretation (items 1, 2, 3), Analysis (items 4, 5, 6), Inference (items 10, 11, 12), and Evaluation (items 7, 8, 9). Item development followed a standard blueprint-based procedure, encompassing the construction of a blueprint based on critical thinking indicators, item writing, qualitative review, and empirical trial.

Content validity of the instrument was established through expert judgment assessment by lecturers specialising in mathematics education, using Aiken's V index (Aiken, 1985). Based on the calculations, a mean Aiken's V value of 0.80 was obtained, classified in the high validity category. This result indicates that the mathematical critical thinking skills test instrument is content-valid.

The selection of these four indicators was based on several considerations. First, the four indicators received the highest consensus in Facione's (2015) Delphi study and are most relevant to the context of mathematical problem-solving in elementary school. Second, the Explanation indicator was not used because it requires open-ended written responses incompatible with the dichotomous assessment format. Third, the Self-regulation indicator was not used because it is a metacognitive process that occurs continuously and therefore cannot be validly measured through items with a single response (Dwyer, Hogan, & Stewart, 2014). Table 1 summarises the four critical thinking indicators used, their operational definitions, and the corresponding items.

Table 1. Research Instrument Summary

No.	Indicator	Item
1	Interpretation Students can understand and explain the meaning of information in a mathematical problem context logically and relevantly.	Items 1, 2, 3 — contextual word problems involving palm oil harvest, library books, and classroom chairs. Students identify known information, unknown quantities, and construct mathematical sentences.
2	Analysis Students can identify, break down, and connect important information in a	Items 4, 5, 6 — word problems involving school canteen stock, study tour buses, and leaf collection. Students

ISSN: 2797-3174

No.	Indicator	Item
	mathematical problem logically to determine the appropriate solution strategy.	explain solution steps and construct mathematical sentences.
3	Inference Students can draw logical conclusions based on evidence and calculation results obtained from a mathematical problem.	Items 10, 11, 12 — problems involving recycling bottle collection, paper stock for LKPD, and sports competition water supply. Students calculate totals, assess sufficiency, and provide reasoned conclusions.
4	Evaluation Students can assess the relevance and correctness of information in a mathematical problem.	Items 7, 8, 9 — problems where two characters give different answers. Students verify each claim, identify who is correct, and explain the calculation error.

2.5. Data Analysis Technique

All data analysis was conducted using Winsteps version 3.73. Student responses were entered as a 49 (persons) × 12 (items) matrix of dichotomous scores (1 = correct, 0 = incorrect) and analysed with the Rasch dichotomous model, in which the probability of a correct response is $P(X=1|\theta, \delta) = \exp(\theta-\delta) / [1 + \exp(\theta-\delta)]$, where θ denotes person ability and δ denotes item difficulty on a shared logit scale. Item and person parameters were estimated using Winsteps' default joint maximum likelihood estimation (JMLE) procedure, and no missing responses were present in the dataset. Analysis covered four psychometric dimensions examined simultaneously. First, construct validity: the fit of items to the Rasch model was assessed through three complementary statistics: Outfit Mean Square (MNSQ) with an acceptance range of 0.5–1.5; Outfit Z-Standard (ZSTD) with an acceptance range of ±2.0; and Point Measure Correlation (Pt Measure Corr) with an acceptance range of 0.40–0.85. An item was declared valid and retained if it met at least two of the three criteria, with priority given to MNSQ and Pt Measure Corr (Bond, Yan, & Heene, 2021; Sumintono & Widhiarso, 2015).

Second, reliability. Internal consistency was measured using KR-20 (Kuder–Richardson Formula 20; Kuder & Richardson, 1937) with the following classification: < 0.50 poor; 0.50–0.60 moderate; 0.60–0.70 acceptable; 0.70–0.80 good; and > 0.80 very good. Person and item reliability and separation indices were evaluated using the criteria of Sumintono and Widhiarso (2015): < 0.67 weak; 0.67–0.80 adequate; 0.80–0.90 good; 0.91–0.94 very good; and > 0.94 excellent.

Third, item discrimination. Item discrimination was analysed based on the Pt Measure Corr classification of Sumintono and Widhiarso (2015): < 0.20 very weak; 0.20–0.39 weak; 0.40–0.59 adequate; 0.60–0.79 good; and ≥ 0.80 very good. Fourth, item difficulty level. Item difficulty is expressed in logit measures with the following classification: > 1.82 very difficult; 0.91–1.82 difficult; −0.91 to +0.91 moderate; −1.82 to −0.91 easy; and < −1.82 very easy (Soeharto & Csapó, 2022). Additionally, the Wright Person–Item Map was used to visualise the alignment between the distribution of student abilities and the item difficulty hierarchy on a shared logit scale (Zeniarja, Utami, & Irawan, 2023). The complete analytical procedure, from data entry through the four psychometric dimensions, is summarised in Figure 1.

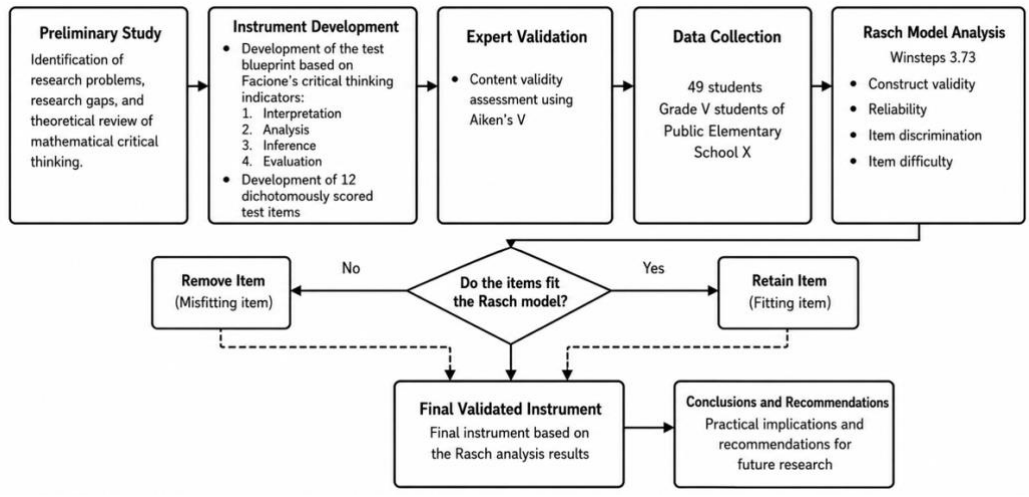


Figure 1. Research Flowchart

ISSN: 2797-3174

3. Results and Discussion

3.1. Results

3.1.1. Construct Validity: Item Fit Statistics

Construct validity of the instrument was tested using the Rasch model. This analysis aimed to determine the extent to which student responses to each item conformed to the expected measurement model, referring to three item fit criteria: Outfit MNSQ in the range 0.5–1.5, Outfit ZSTD in the range −2.0 to +2.0, and Pt Measure Corr with a critical minimum value of 0.40 (Sumintono & Widhiarso, 2015). An item was declared valid if it met all criteria simultaneously. Table 2 presents the item fit statistics and validity decision for each of the 12 items.

Table 2. Item Fit Statistics and Validity Decisions

Item	Outfit MNSQ	Outfit ZSTD	Pt Measure Corr	Decision
Item 1	0.66	−1.1	0.22	Not Valid
Item 2	0.96	−0.1	0.51	Valid ✓
Item 3	1.12	0.6	0.52	Valid ✓
Item 4	0.65	−1.7	0.72	Valid ✓
Item 5	0.59	−2.3	0.66	Valid ✓
Item 6	0.96	−0.1	0.60	Valid ✓
Item 7	1.15	0.7	0.61	Valid ✓
Item 8	1.52	2.3	0.47	Valid ✓
Item 9	1.44	1.9	0.52	Valid ✓
Item 10	1.00	0.1	0.67	Valid ✓
Item 11	0.56	−2.3	0.69	Valid ✓
Item 12	0.73	−0.6	0.21	Not Valid

Of 12 items, 10 items (83.3%) met the model fit criteria (Items 2–11), while 2 items (Items 1 and 12) were classified as invalid. Item 8 (MNSQ = 1.52) approached the upper MNSQ boundary but met the ZSTD and Pt Measure Corr criteria and was therefore retained. Items 1 and 12 failed to exceed the critical Pt Measure Corr threshold (0.22 and 0.21 respectively, both < 0.40) and failed on at least one additional criterion, confirming their exclusion from the final instrument. Figure 2 presents the Item Statistics Misfit Order output of the Rasch model.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
8	110	49	-.31	.12	1.64	3.2	1.52	2.3	A .47	.58	26.5	28.4	Soal 8
9	81	49	.12	.12	1.61	3.0	1.44	1.9	B .52	.54	20.4	30.7	Soal 9
7	76	49	.20	.12	1.25	1.4	1.15	.7	C .61	.53	24.5	31.4	Soal 7
3	124	49	-.52	.13	1.06	.4	1.12	.6	D .52	.60	32.7	27.8	Soal 3
10	109	49	-.29	.12	1.12	.7	1.00	.1	E .67	.58	24.5	28.4	Soal 10
2	126	49	-.56	.13	.75	-1.4	.96	-.1	F .51	.60	24.5	27.7	Soal 2
6	116	49	-.40	.12	.95	-.2	.96	-.1	f .60	.59	22.4	28.0	Soal 6
12	17	49	1.60	.22	.63	-1.1	.73	-.6	e .21	.29	61.2	68.7	Soal 12
1	34	49	1.00	.16	.52	-2.3	.66	-1.1	d .22	.39	46.9	44.3	Soal 1
4	130	49	-.62	.13	.65	-2.1	.65	-1.7	c .72	.60	28.6	28.6	Soal 4
5	119	49	-.44	.12	.58	-2.7	.59	-2.3	b .66	.59	44.9	26.7	Soal 5
11	74	49	.23	.12	.56	-2.9	.56	-2.3	a .69	.52	34.7	31.8	Soal 11
MEAN	93.0	49.0	.00	.14	.94	-.3	.95	-.2			32.7	33.5	
S.D.	35.6	.0	.66	.03	.38	2.0	.31	1.4			11.8	11.5	

Figure 2. Output: Item Statistics Misfit Order Rasch Model

3.1.2. Reliability Analysis

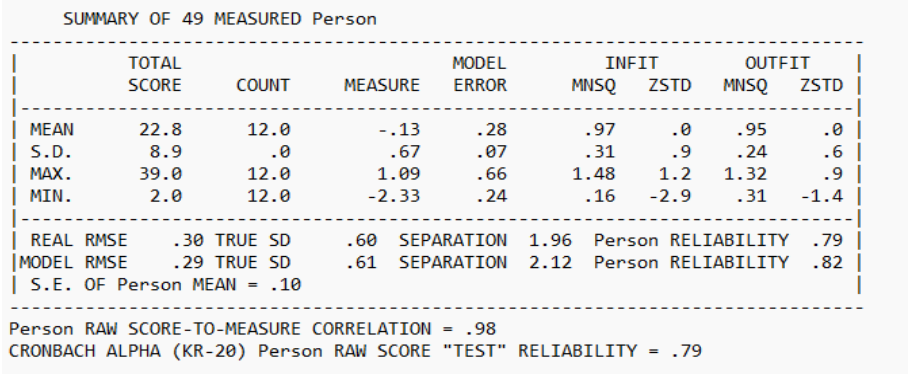
Instrument reliability was evaluated through three indicators: internal consistency (KR-20), person reliability, and item reliability, with interpretation categories referenced to Sumintono and Widhiarso (2015). Table 3 presents the value, category, and interpretation of each reliability indicator.

Table 3. Reliability Indicators

Reliability Indicator	Value	Category	Interpretation
Cronbach's Alpha (KR-20)	0.79	Good	Adequate internal consistency
Person Reliability (Real)	0.79	Adequate	Consistent respondent measurement
Person Reliability (Model)	0.82	Good	Adequate ability differentiation
Item Reliability (Real)	0.95	Excellent	Very high item quality stability
Item Reliability (Model)	0.95	Excellent	Very consistent item difficulty hierarchy

ISSN: 2797-3174

A KR-20 value of 0.79 (Good category) confirms adequate internal consistency. Person reliability (0.79–0.82) indicates that the instrument can reliably differentiate respondent ability, while item reliability of 0.95 (Excellent) indicates a very stable and replicable item difficulty hierarchy (Sumintono & Widhiarso, 2015). Figure 3 presents the corresponding Winsteps person summary statistics output.



SUMMARY OF 49 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	22.8	12.0	-.13	.28	.97	.0	.95	.0
S.D.	8.9	.0	.67	.07	.31	.9	.24	.6
MAX.	39.0	12.0	1.09	.66	1.48	1.2	1.32	.9
MIN.	2.0	12.0	-2.33	.24	.16	-2.9	.31	-1.4

REAL RMSE	.30	TRUE SD	.60	SEPARATION	1.96	Person RELIABILITY .79
MODEL RMSE	.29	TRUE SD	.61	SEPARATION	2.12	Person RELIABILITY .82
S.E. OF Person MEAN = .10						

Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .79

Figure 3. Output: Summary Statistics — Person (Rasch Model)

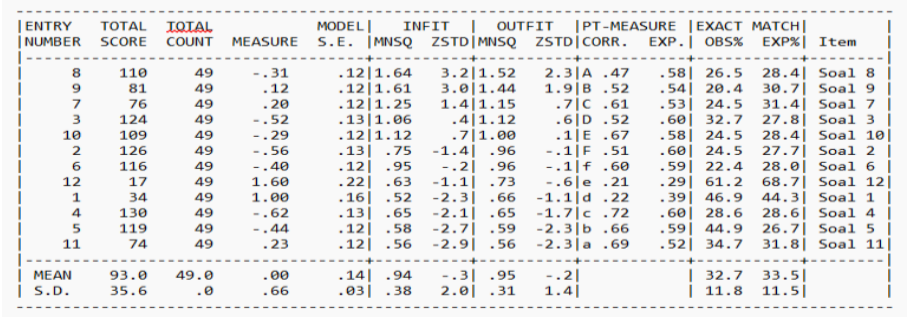
3.1.3. Item Discrimination

Discrimination analysis aims to determine the extent to which each item can differentiate between high- and low-ability persons. In the Rasch model approach, discrimination is assessed through the Point-Measure Correlation (Pt Measure Corr) value in the Winsteps output: the higher the value, the better the item discriminates ability. Classification criteria used: ≥0.60 = Good, 0.40–0.59 = Adequate, and < 0.40 = Weak. Table 4 presents the discrimination power, status, and retention decision for each item.

Table 4. Item Discrimination Power

Item	Pt Measure Corr	Discrimination Category	Status	Decision
Item 4	0.72	Good	Adequate	Retained
Item 5	0.66	Good	Adequate	Retained
Item 7	0.61	Good	Adequate	Retained
Item 10	0.67	Good	Adequate	Retained
Item 11	0.69	Good	Adequate	Retained
Item 6	0.60	Good	Adequate	Retained
Item 2	0.51	Adequate	Adequate	Retained
Item 3	0.52	Adequate	Adequate	Retained
Item 8	0.47	Adequate	Adequate	Retained
Item 9	0.52	Adequate	Adequate	Retained
Item 1	0.22	Weak	Inadequate	Excluded
Item 12	0.21	Weak	Inadequate	Excluded

Of 12 items analysed, 5 items (Items 4, 5, 6, 7, 10, 11) showed Good discrimination (Pt Measure Corr 0.60–0.72), 4 items (Items 2, 3, 8, 9) showed Adequate discrimination (0.47–0.60), while 2 items (Items 1 and 12) showed Weak discrimination (0.21–0.22). Accordingly, 10 items were declared adequate and retained, while Items 1 and 12 were declared inadequate and recommended for removal due to insufficient ability to discriminate person ability effectively. Figure 4 presents the corresponding Winsteps item misfit output.



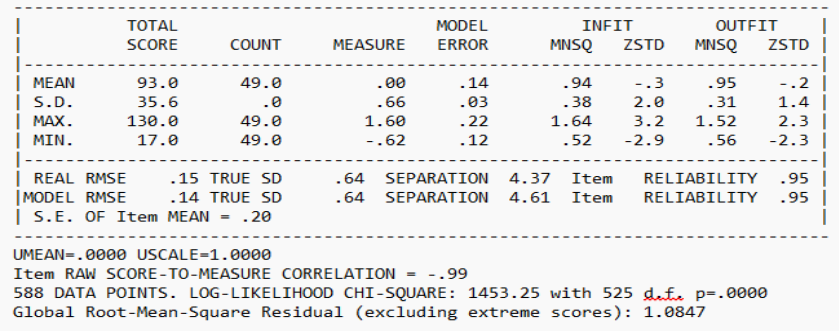
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
8	110	49	-.31	.12	1.64	3.2	1.52	2.3	A .47	.58	26.5	28.4	Soal 8
9	81	49	.12	.12	1.61	3.0	1.44	1.9	B .52	.54	20.4	30.7	Soal 9
7	76	49	.20	.12	1.25	1.4	1.15	.7	C .61	.53	24.5	31.4	Soal 7
3	124	49	-.52	.13	1.06	.4	1.12	.6	D .52	.60	32.7	27.8	Soal 3
10	109	49	-.29	.12	1.12	.7	1.00	.1	E .67	.58	24.5	28.4	Soal 10
2	126	49	-.56	.13	.75	-1.4	.96	-.1	F .51	.60	24.5	27.7	Soal 2
6	116	49	-.40	.12	.95	-.2	.96	-.1	f .60	.59	22.4	28.0	Soal 6
12	17	49	1.60	.22	.63	-1.1	.73	-.6	e .21	.29	61.2	68.7	Soal 12
1	34	49	1.00	.16	.52	-2.3	.66	-1.1	d .22	.39	46.9	44.3	Soal 1
4	130	49	-.62	.13	.65	-2.1	.65	-1.7	c .72	.60	28.6	28.6	Soal 4
5	119	49	-.44	.12	.58	-2.7	.59	-2.3	b .66	.59	44.9	26.7	Soal 5
11	74	49	.23	.12	.56	-2.9	.56	-2.3	a .69	.52	34.7	31.8	Soal 11
MEAN	93.0	49.0	.00	.14	.94	-.3	.95	-.2			32.7	33.5	
S.D.	35.6	.0	.66	.03	.38	2.0	.31	1.4			11.8	11.5	

Figure 4. Output: Item Statistics Misfit Order — Rasch Model

Figure 5 presents summary statistics of the 12 items, with a mean difficulty level of 0.00 logits (range –0.62 to 1.60). INFIT (0.94) and OUTFIT (0.95) values close to 1.0 indicate that items fit the model. Item Reliability is

ISSN: 2797-3174

very high (0.95) and Item Separation is high (4.37–4.61), indicating that items can clearly and consistently differentiate difficulty levels. Overall, item quality is classified as good.



	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	93.0	49.0	.00	.14	.94	-.3	.95	-.2
S.D.	35.6	.0	.66	.03	.38	2.0	.31	1.4
MAX.	130.0	49.0	1.60	.22	1.64	3.2	1.52	2.3
MIN.	17.0	49.0	-.62	.12	.52	-2.9	.56	-2.3

REAL RMSE .15 TRUE SD .64 SEPARATION 4.37 Item RELIABILITY .95
MODEL RMSE .14 TRUE SD .64 SEPARATION 4.61 Item RELIABILITY .95
S.E. OF Item MEAN = .20

UMEAN=.0000 USCALE=1.0000
Item RAW SCORE-TO-MEASURE CORRELATION = -.99
588 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 1453.25 with 525 d.f. p=.0000
Global Root-Mean-Square Residual (excluding extreme scores): 1.0847

Figure 5. Output: Summary Statistics — Item (Rasch Model)

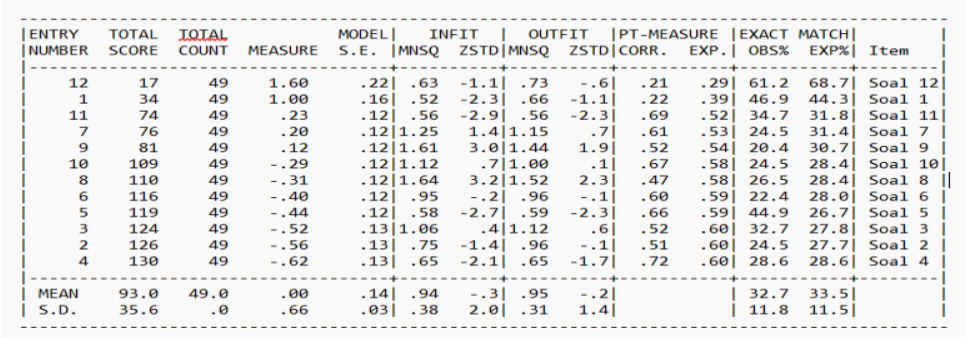
3.1.4. Item Difficulty Analysis and Wright Person–Item Map

Item difficulty analysis aims to determine the position of each item on the logit scale of the Rasch model, reflecting how difficult the item is to answer correctly. Higher measure values indicate more difficult items, while lower values indicate easier items. Difficulty categories are determined based on the mean and standard deviation (SD) of item measures: Very Difficult (> 1.82 logits), Difficult (0.23–1.82), Moderate (−0.91 to +0.91), and Easy (< −0.91). Table 5 presents the measure, difficulty level, validity, and decision for each item.

Table 5. Item Difficulty Levels

Item	Measure (Logit)	Difficulty Level	Validity	Decision
Item 12	1.60	Difficult	Not Valid	Excluded
Item 1	1.00	Difficult	Not Valid	Excluded
Item 11	0.23	Moderate	Valid	Retained
Item 7	0.20	Moderate	Valid	Retained
Item 9	0.12	Moderate	Valid	Retained
Item 10	−0.29	Moderate	Valid	Retained
Item 8	−0.31	Moderate	Valid	Retained
Item 6	−0.40	Moderate	Valid	Retained
Item 5	−0.44	Moderate	Valid	Retained
Item 3	−0.52	Moderate	Valid	Retained
Item 2	−0.59	Moderate	Valid	Retained
Item 4	−0.66	Moderate	Valid	Retained

Based on Table 5, no items fall in the Very Difficult or Easy categories. Items 1 and 12 are in the Difficult category (measures 1.00 and 1.60 logits) but are both invalid (misfit) and are therefore recommended for removal. The remaining 10 items (Items 2, 3, 4, 5, 6, 7, 8, 9, 10, 11) fall in the Moderate category (range −0.66 to +0.23 logits) and are valid and are therefore retained. These results indicate that the majority of items have balanced difficulty levels appropriate for measuring person ability in the moderate ability range. Figure 6 presents the corresponding Winsteps item summary statistics output.



ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
12	17	49	1.60	.22	.63	-1.1	.73	-.6	.21	.29	61.2	68.7	Soal 12
1	34	49	1.00	.16	.52	-2.3	.66	-1.1	.22	.39	46.9	44.3	Soal 1
11	74	49	.23	.12	.56	-2.9	.56	-2.3	.69	.52	34.7	31.8	Soal 11
7	76	49	.20	.12	1.25	1.4	1.15	.7	.61	.53	24.5	31.4	Soal 7
9	81	49	.12	.12	1.61	3.0	1.44	1.9	.52	.54	20.4	30.7	Soal 9
10	109	49	-.29	.12	1.12	.7	1.00	.1	.67	.58	24.5	28.4	Soal 10
8	110	49	-.31	.12	1.64	3.2	1.52	2.3	.47	.58	26.5	28.4	Soal 8
6	116	49	-.40	.12	.95	-.2	.96	-.1	.60	.59	22.4	28.0	Soal 6
5	119	49	-.44	.12	.58	-2.7	.59	-2.3	.66	.59	44.9	26.7	Soal 5
3	124	49	-.52	.13	1.06	.4	1.12	.6	.52	.60	32.7	27.8	Soal 3
2	126	49	-.56	.13	.75	-1.4	.96	-.1	.51	.60	24.5	27.7	Soal 2
4	130	49	-.62	.13	.65	-2.1	.65	-1.7	.72	.60	28.6	28.6	Soal 4
MEAN	93.0	49.0	.00	.14	.94	-.3	.95	-.2			32.7	33.5	
S.D.	35.6	.0	.66	.03	.38	2.0	.31	1.4			11.8	11.5	

Figure 6. Output: Summary Statistics — Item (Rasch Model)

The Wright Person–Item Map places persons (ability) and items (difficulty) on the same logit scale, enabling direct comparison between the distribution of participant ability and the item difficulty hierarchy. Persons are displayed on the left side (X symbol) while items are displayed on the right side of the map.

ISSN: 2797-3174

Based on Figure 7, the map shows that Item 12 is the most difficult item (positioned at the top, around logit +2), while Item 4 is the easiest (positioned at the bottom, around logit −0.6). Most items (Items 2, 3, 4, 5, 6, 7, 8, 9, 10, and 11) cluster around logit 0, aligned with the distribution of most persons, who are also concentrated in the range of logit −1 to +1. This indicates that item difficulty is reasonably well suited to the average person ability, although there is a gap at the upper part of the map: Items 1 and 12 (Difficult category) have few persons with commensurate ability, confirming the earlier finding that both items are misfitting and should be removed. Conversely, no items represent the very low ability area (below logit −1), indicating an absence of easy items for measuring low-ability persons in greater detail.

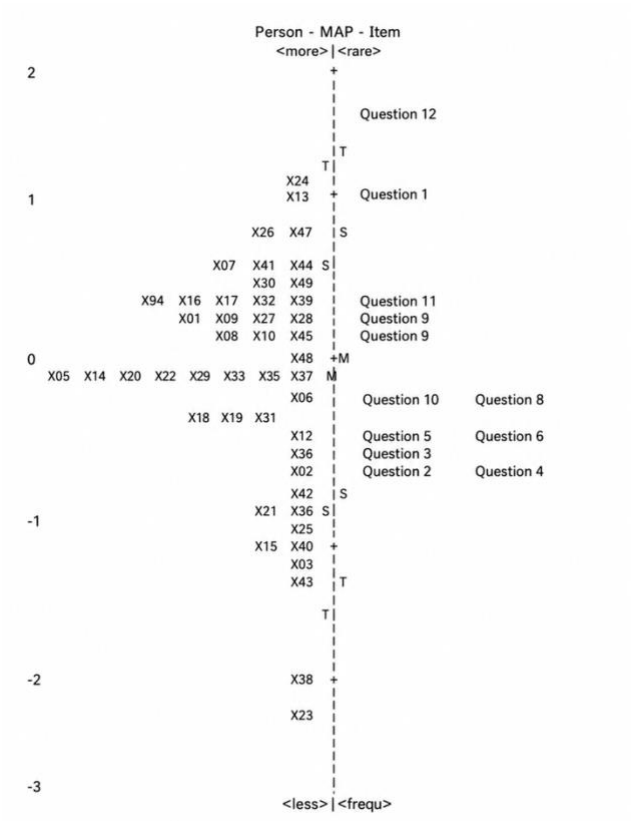


Figure 7. Wright Person–Item Map (Rasch Model)

3.2. Discussion

3.2.1. Construct Validity and Item Fit (RQ1)

The first research question (RQ1) aimed to identify which items met Rasch model fit criteria. Analysis results showed that 10 of 12 items (83.3%) met the fit criteria, while Items 1 and 12 were declared invalid and must be excluded from the instrument. This 83.3% retention rate reflects an item construction process in which the blueprint design and expert content-validation stage were closely aligned with the Facione (2015) framework prior to empirical testing and is consistent with international Rasch-based validation practice, where items showing inadequate fit statistics or negative point-measure correlations are typically identified and removed at this stage (Santos et al., 2016).

The exclusion of Items 1 and 12 has a significant constructive impact on instrument quality. Item 1 measures the Interpretation indicator (C2 — Understanding) while Item 12 measures the Inference indicator (C4 — Analysing). The failure pattern of both items, low Pt Measure Corr (0.21–0.22) combined with high item difficulty (logit +1.00 and +1.60)—indicates a psychometric paradox: these items are difficult to answer correctly, yet fail to differentiate high-ability from low-ability students. This phenomenon is known in the Rasch literature as misfitting items with high difficulty, theoretically explained by Bond, Yan, and Heene (2021) as a consequence of ambiguous item wording that causes students to respond randomly without a pattern predictable by the model. Tesio and Scarano (2024) explain that a pattern of low Outfit MNSQ combined with low Pt Measure Corr indicates items that are too easily guessable for some students yet ambiguous for others, not merely difficult items. The concrete impact of excluding both items is an improvement in the construct coherence of the instrument: the remaining 10 items form a cleaner logit continuum from −0.66 to +0.23, with more even coverage across the four critical thinking indicators of Facione (2015): Interpretation (Items 2, 3), Analysis (Items 4, 5, 6), Evaluation (Items 7, 8, 9), and Inference (Items 10, 11).

ISSN: 2797-3174

Item 8 is a borderline case warranting special attention. With MNSQ = 1.52 and ZSTD = 2.3, this item is precisely at the upper boundary of the fit criteria, yet its Pt Measure Corr remains adequate (0.47). The decision to retain Item 8 follows the recommendation of Boone and Noltemeyer (2017) that when two of the three fit criteria are met with Pt Measure Corr as one of them, the item still provides meaningful discriminative contribution. The underfit pattern of Item 8 is consistent with the findings of Sujatmika et al. (2025), who found that Evaluation items requiring step-by-step evidence assessment tend to generate greater response variability due to divergent student assessment standards; such misfit is typically attributed to ambiguous item wording that prompts guessing rather than reasoned response (Santos et al., 2016). In practice, Item 8 should be revised in its instruction wording to make the response strategy more focused, so that its MNSQ value can be reduced to the optimal range of 0.8–1.2 in subsequent trials.

3.2.2. Internal Consistency and Reliability (RQ2)

The second research question (RQ2) targeted the level of internal consistency, person reliability, and item reliability of the instrument. A KR-20 value of 0.79 (Good category) falls within the range generally considered adequate for formative assessment and instructional research purposes for an instrument of this length and sample size (Sumintono & Widhiarso, 2015). This is broadly comparable to KR-20 and Person Separation Reliability coefficients reported for other elementary-level Rasch-calibrated instruments internationally, which have similarly fallen in the moderate-to-good range (KR-20 ≈ 0.72–0.85; Person Separation Reliability ≈ 0.69–0.78) despite covering a different construct (Santos et al., 2016).

Person Reliability (Real = 0.79; Model = 0.82) is in the Adequate–Good range. This value indicates that the instrument can consistently differentiate respondents based on their critical thinking ability, although there is room for improvement by adding items at higher difficulty levels. Person reliability values correlate with sample size, larger samples provide more stable estimates (Sumintono & Widhiarso, 2015). Accordingly, Person Reliability of 0.79 at n = 49 is an adequate achievement for an instrument trial sample at the elementary school level.

The most prominent finding in the reliability dimension is Item Reliability = 0.95 (Excellent) with Item Separation = 4.37–4.61. This means that the difficulty hierarchy of the 10 retained items is very stable and can be replicated consistently across similar samples, comparable to the very high Item Separation Reliability coefficients (0.94–0.96) reported for Rasch-calibrated elementary-level instruments internationally (Santos et al., 2016). The practical implication of Item Separation 4.37–4.61 is that the instrument can distinguish at least four to five statistically different item difficulty groups on the logit scale, which is an important psychometric prerequisite for instruments designed to measure hierarchical constructs such as Facione's (2015) critical thinking (Bond, Yan, & Heene, 2021).

Person Separation (1.96–2.12) confirms that the instrument can differentiate two statistically distinct critical thinking ability strata. Although lower than Item Separation, this is common for instruments with a limited number of items (n = 10 items) and small sample sizes (Sumintono & Widhiarso, 2015). The implication for future instrument development is the need to add two to three items with difficulty above +0.50 logits so that the instrument can differentiate three student ability strata: low, moderate, and high.

3.2.3. Item Discrimination (RQ3)

The third research question (RQ3) examined the discriminating power of each item. Analysis results show a pattern of discrimination that varies systematically in accordance with the hierarchy of Facione's (2015) critical thinking skills. Five items achieved the Good category (Pt Measure Corr = 0.60–0.72): Items 4 and 5 (Analysis), Item 7 (Evaluation), and Items 10 and 11 (Inference). Five items were in the Adequate category (Pt Measure Corr = 0.47–0.60): Items 2 and 3 (Interpretation), Item 6 (Analysis), and Items 8 and 9 (Evaluation). No retained items fell below the Adequate threshold. This distribution indicates that all 10 retained items provide meaningful discriminative contribution in measuring students' mathematical critical thinking ability.

The pattern whereby Analysis items (Items 4, 5) and Inference items (Items 10, 11) achieved the highest discrimination (0.66–0.72) is consistent with Facione's (2015) characterisation that Analysis and Inference require multi-step reasoning demanding deeper cognitive engagement than mere comprehension. Sujatmika et al. (2025) found a similar pattern: items requiring multi-step reasoning and conclusion-drawing from available evidence consistently produced higher Pt Measure Corr than items requiring only fact recognition.

Conversely, Interpretation items (Items 2 and 3) only achieved the Adequate category (0.51–0.52). This is consistent with the nature of the Interpretation indicator, which requires meaning comprehension without inferential reasoning, making correct answers more likely by chance without deep understanding. The exclusion of Item 1 (Interpretation, Pt Measure Corr = 0.22) and Item 12 (Inference, Pt Measure Corr = 0.21) directly impacted an increase in the average discrimination of the final instrument: from a mean of 0.54 across 12 items to 0.61 across the 10 retained items, a psychometrically meaningful improvement as it crosses the boundary from the Adequate to the Good category (Sumintono & Widhiarso, 2015).

ISSN: 2797-3174

3.2.4. Item Difficulty Distribution and Wright Person–Item Map (RQ4)

The fourth research question (RQ4) focused on the distribution of items across the difficulty hierarchy. All 10 retained items fall in the Moderate category (range −0.66 to +0.23 logits); none entered the Easy (< −0.91) or Difficult (> +0.91) categories. This concentration of items in the Moderate range is psychometrically advantageous as it maximises measurement precision for the majority of students with average ability in the sampled population (Soeharto & Csapó, 2022), though it represents a trade-off, since it limits discrimination for students at the extreme ends of the ability distribution.

The Wright Person–Item Map confirms near-optimal targeting of the instrument: mean person ability (M = 0.0 logits) coincides with mean item difficulty (M = 0.0 logits), which is the ideal targeting condition (Bond, Yan, & Heene, 2021), consistent with the construct-modelling approach demonstrated by Rittle-Johnson, Matthews, Taylor, and McEldoon (2011) for diagnosing mathematics knowledge via a Wright map. The optimal targeting in this study indicates that the item development process successfully calibrated the cognitive demands of items to match the ability characteristics of Grade V students at public elementary school X.

The difficulty hierarchy based on Facione's (2015) critical thinking indicators receives partial empirical support from the Wright Map. Inference items (Items 10, 11; logits −0.29 to +0.23) and Evaluation items (Items 7, 8, 9; logits −0.44 to +0.20) occupied the upper positions in the retained item difficulty hierarchy, consistent with the theoretical prediction that higher-order critical thinking skills produce more difficult items. Interpretation items (Items 2, 3; logits −0.52 to −0.59) occupied the easiest positions, in line with their lower cognitive demands (C2). Analysis items (Items 4, 5, 6; logits −0.29 to −0.66) occupied the lower-middle position. These findings are broadly consistent with Sujatmika et al. (2025), whose polytomous Rasch analysis likewise found Inference and Evaluation items more difficult than Analysis and Interpretation items in a critical thinking instrument, despite differences in context and scoring model.

The gap visible in the Wright Map between the highest retained item (Item 11, Inference, +0.23 logits) and the excluded items (Item 1 Interpretation +1.00 logits; Item 12 Inference +1.60 logits) exposes a coverage deficit in the difficulty range from +0.24 to +0.99 logits. This deficit means the instrument cannot precisely measure the critical thinking ability of above-average students, particularly on high-level Inference indicators. This pattern is consistent with the Person Separation obtained (1.96–2.12), which is lower than Item Separation (4.37–4.61). To address this deficit, future instrument development needs to add two to three new items targeting the +0.40 to +1.20 logit range on Inference and Evaluation indicators with multi-step problem contexts.

3.2.5. Overall Implications of Rasch Analysis

Overall, the Rasch-based validation process produced a qualitative transformation of the instrument: from 12 items covering all four Facione (2015) critical thinking indicators with two items of low psychometric quality, to 10 items with homogeneous construct quality, excellent reliability, and an empirically documented difficulty hierarchy. This transformation is not merely a reduction in item count but a fundamental improvement in construct validity. The Rasch model enables researchers not only to detect poor items but also to diagnose the cause of their failure, information unavailable through conventional CTT approaches (Guler & Uyanik, 2021; Hambleton & Jones, 1993).

From a practical field perspective, the validated 10-item instrument provides three concrete benefits for teachers at public elementary school X and similar elementary schools. First, the Wright Person–Item Map functions as an individual diagnostic map: students with ability estimates below −0.30 logits, challenged by almost all items—require explicit critical thinking scaffolding interventions, such as direct teacher modelling of analysis and inference strategies; a training programme built around structured mathematical problem-solving strategies has been shown to produce measurable gains in critical thinking among grade-school students, supporting the feasibility of this recommendation (Alfayez, Aladwan, & Shaheen, 2022). Second, students with ability above +0.20 logits are not sufficiently challenged by this instrument and require enrichment with higher-difficulty items or more complex critical thinking project tasks. Third, the difference in discrimination profiles between stronger indicators (Analysis and Inference with Good discrimination) and relatively weaker ones (Interpretation and Evaluation with Adequate discrimination) provides clear direction for learning redesign: the portion of learning developing Interpretation and Evaluation skills needs to be strengthened, particularly through activities involving reading mathematical data representations and critically evaluating evidence.

For future research, three developmental directions are recommended. First, a confirmatory re-validation study with a larger sample (n ≥ 100) from multiple schools in the same region should be conducted to test the generalisability of item parameters. Second, Differential Item Functioning (DIF) analysis should be included to examine whether items function differently across gender or socio-economic background subgroups. Third, a longitudinal design integrating test-retest reliability will establish the temporal stability of the instrument, a prerequisite for its use in monitoring the development of student critical thinking over time.

ISSN: 2797-3174

3.3. Implications

The validated 10-item instrument has both theoretical and practical implications. Theoretically, this study demonstrates that a four-dimensional simultaneous Rasch analysis produces more informative psychometric evidence than the partial approaches commonly reported in Indonesian measurement literature, and provides independent empirical support for the Facione (2015) critical thinking skill hierarchy in an elementary school mathematics context. Practically, the Wright Person–Item Map enables individualised diagnostic profiling: teachers can identify students requiring critical thinking scaffolding (ability below −0.30 logits) and those needing enrichment tasks (above +0.20 logits), while the discrimination profile directs instructional attention toward strengthening Evaluation and Interpretation sub-skills. The instrument is recommended for formative assessment in Grade V classrooms implementing the *Kurikulum Merdeka*.

These findings also speak to a broader body of research on critical thinking in mathematics education. Because critical thinking and academic achievement have been shown to reinforce one another bidirectionally from the upper-elementary years onward (Lv, Zhou, & Ren, 2025), an instrument capable of diagnosing specific critical thinking weaknesses at Grade V, as the Wright Person–Item Map developed here does, offers a concrete entry point for interrupting low-achievement trajectories before they compound in later schooling. This is consistent with the broader systematic-review evidence that intervention research on critical thinking in mathematics remains hampered by assessment tools that are not clearly grounded in an explicit critical thinking framework (Wang & Abdullah, 2024); the four-dimensional, Facione-grounded design adopted in this study directly addresses that gap for the elementary mathematics context.

3.4. Limitations

This study has several limitations. First, the sample of 49 students from a single school limits the generalisability of item parameters; confirmatory re-validation with a larger, multi-school sample ($n \geq 100$) is needed. Second, all retained items cluster within the Moderate difficulty range (−0.66 to +0.23 logits), leaving a coverage gap for above-average students and contributing to the relatively lower Person Separation (1.96–2.12). Third, the cross-sectional single-context design means temporal stability and cross-context applicability of item parameters have not been established. Finally, two of the six Facione (2015) sub-skills (Explanation and Self-Regulation) were excluded, and Differential Item Functioning (DIF) analysis was not conducted, limiting claims about measurement equity across student subgroups; Rasch-based DIF techniques capable of detecting such bias are well established and represent a natural next step for future validation of this instrument (Wyse & Mapuranga, 2009).

4. Conclusion

This study conducted a comprehensive four-dimensional Rasch-based psychometric evaluation of a 12-item mathematical critical thinking test instrument grounded in the Facione (2015) framework of Interpretation, Analysis, Evaluation, and Inference, administered to 49 elementary school students. Ten of 12 items (83.3%) met the Rasch model fit criteria and were retained (RQ1); the instrument showed good internal consistency and reliability, with KR-20 = 0.79, Item Reliability = 0.95 (Excellent), and Person Reliability = 0.79–0.82 (RQ2); the 10 retained items exhibited Adequate-to-Good discrimination (Pt Measure Corr = 0.47–0.72) (RQ3); and all retained items fell within the Moderate difficulty range (logit −0.66 to +0.23), with the Wright Person–Item Map confirming near-optimal targeting of item difficulty to the sampled students' ability range and partial empirical support for the Facione (2015) difficulty hierarchy-Inference and Evaluation items at the upper range, Analysis items in the middle position, and Interpretation items at the easiest positions, while the two excluded items (Item 1 Interpretation and Item 12 Inference) sat at the upper extreme as invalid, difficult items (RQ4). Taken together, these findings indicate that the four-dimensional Rasch approach yielded a psychometrically sound 10-item instrument, recommended for formative assessment of elementary school students' mathematical critical thinking, with future development needed to add higher-difficulty Inference and Evaluation items (+0.50 to +1.50 logits) to close the identified coverage gap.

Author Contributions

Yogina Pratiwi: Conceptualization, Methodology, Investigation, Validation, Writing – Original Draft. Prihantini: Supervision, Methodology, Writing – Review & Editing. Abubakar: Formal Analysis, Data Curation, Writing – Review & Editing.

Funding

No funding support was received.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ISSN: 2797-3174

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration on AI Use

The authors declare that AI-assisted tools were used only to improve readability and language under strict human oversight; no content, ideas, analyses, or conclusions were generated by AI.

References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement, 45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Alfayez, M. Q. E., Aladwan, S. Q. A., & Shaheen, H. R. A. (2022). The effect of a training program based on mathematical problem-solving strategies on critical thinking among seventh-grade students. *Frontiers in Education, 7*, 870524. https://doi.org/10.3389/feduc.2022.870524

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education, 4*(1), 1416898. https://doi.org/10.1080/2331186X.2017.1416898

Chang, I. (2023). Early numeracy and literacy skills and their influences on fourth-grade mathematics achievement: A moderated mediation model. *Large-Scale Assessments in Education, 11*(1), 18. https://doi.org/10.1186/s40536-023-00168-6

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity, 12*, 43–52. https://doi.org/10.1016/j.tsc.2014.02.002

Facione, P. A. (2015). *Critical thinking: What it is and why it counts* (2015 update). Insight Assessment.

Guler, N., & Uyanik, G. K. (2021). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Educational Research, 10*(1), 395–407. https://doi.org/10.12973/eu-jer.10.1.395

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. (2022). *Keputusan Kepala BSKAP No. 033/H/KR/2022 tentang capaian pembelajaran pada pendidikan anak usia dini, jenjang pendidikan dasar, dan jenjang pendidikan menengah pada Kurikulum Merdeka*. Kemendikbudristek.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

Lv, X., Zhou, J., & Ren, X. (2025). The bidirectional relationship between critical thinking and academic achievement is independent of general cognitive ability: A three-year longitudinal study on elementary school children. *Learning and Individual Differences, 120*, 102666. https://doi.org/10.1016/j.lindif.2025.102666

Mapeala, R., & Siew, N. M. (2015). The development and validation of a test of science critical thinking for fifth graders. *SpringerPlus, 4*, Article 741. https://doi.org/10.1186/s40064-015-1535-0

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Rittle-Johnson, B., Matthews, P. G., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology, 103*(1), 85–104. https://doi.org/10.1037/a0021334

Santos, S., Cadime, I., Viana, F. L., Prieto, G., Chaves-Sousa, S., Spinillo, A. G., & Ribeiro, I. (2016). An application of the Rasch model to reading comprehension measurement. *Psicologia: Reflexão e Crítica, 29*, Article 38. https://doi.org/10.1186/s41155-016-0044-6

Soeharto, S., & Csapó, B. (2022). Evaluating item difficulty patterns for assessing student problem-solving abilities in science: Insights from Rasch analysis. *Frontiers in Education, 7*, 1021063. https://doi.org/10.3389/feduc.2022.1021063

Sujatmika, S., Sutarno, Masykuri, M., & Prayitno, B. A. (2025). Applying the Rasch model to measure students' critical thinking skills on the science topic of the human circulatory system. *Eurasia Journal of Mathematics, Science and Technology Education, 21*(4), em2622. https://doi.org/10.29333/ejmste/16221

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada asesmen pendidikan*. Trim Komunikata.

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teacher, 35*(1), e838–e848. https://doi.org/10.3109/0142159X.2012.737488

Tesio, L., & Scarano, S. (2024). Interpreting results from Rasch analysis 2: Advanced model applications. *Disability and Rehabilitation, 46*(4), 604–617. https://doi.org/10.1080/09638288.2023.2169772

ISSN: 2797-3174

Wang, Q., & Abdullah, A. H. (2024). Enhancing students' critical thinking through mathematics in higher education: A systemic review. *SAGE Open, 14*(3), 1–15. https://doi.org/10.1177/21582440241275651

Wyse, A. E., & Mapuranga, R. (2009). Differential item functioning analysis using Rasch item information functions. *International Journal of Testing, 9*(4), 333–357. https://doi.org/10.1080/15305050903352040

ISSN: 2797-3174