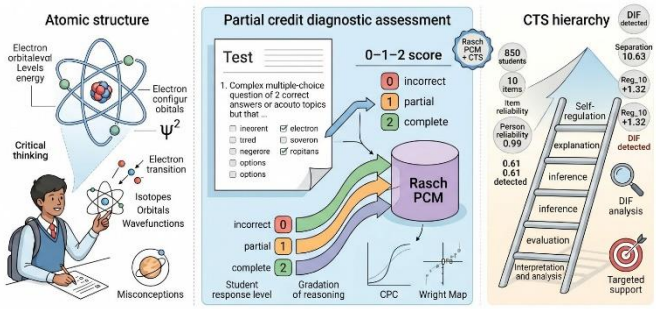
***JKPK (JURNAL KIMIA DAN PENDIDIKAN KIMIA)**, Vol. 11, No. 1, 2026* *pp. 130-146*
Chemistry Education Study Program, Universitas Sebelas Maret *ISSN 2503-4146*
https://jurnal.uns.ac.id/jkpk *ISSN 2503-4154 (online)*

Rasch PCM Diagnostic Analysis of Critical Thinking Item Responses in Indonesian Atomic Structure Learning

Lukman A. R. Laliyo*, Yeyen Katili, Astin Lukum, Akram La Kilo, Masrid Pikoli

Master's Program in Chemistry Education, Postgraduate Program, Universitas Negeri Gorontalo, Gorontalo, Indonesia

ABSTRACT



Measuring critical thinking skills (CTS) in the topic of atomic structure requires a diagnostic instrument capable of capturing students' reasoning patterns accurately, rather than merely distinguishing between correct and incorrect answers. This study aimed to develop and evaluate a complex multiple-choice diagnostic instrument based on the Rasch model to measure students' CTS on the concept of atomic structure. The instrument comprised ten items developed based on Facione's six dimensions of critical thinking: interpretation, analysis, evaluation, inference, explanation, and self-regulation. Data were collected from 850 senior high school students in Gorontalo, Indonesia, and analyzed using the Partial Credit Model (PCM) approach with WINSTEPS 4.5.5 software. Results indicated that item reliability was very high (0.99; separation = 10.63), while person reliability was moderate (0.61; Cronbach's Alpha = 0.65). Infit and Outfit MNSQ values ranged within the ideal threshold (0.99–1.00) with ZSTD values approaching zero, confirming adequate model fit. Category Probability Curve (CPC) analysis confirmed that response categories functioned sequentially, while the Wright Map demonstrated a progressive difficulty hierarchy from the interpretation to the self-regulation dimension. Differential Item Functioning (DIF) analysis revealed differential difficulty levels based on gender and grade level, particularly for self-regulation items, which proved more challenging for female students and Grade XI students. These findings underscore the importance of considering cognitive factors and learning experience in developing Rasch-based CTS diagnostic instruments for chemistry education.

Keywords: *critical thinking skills; atomic structure; diagnostic instrument; Rasch model; partial credit model*
**Corresponding Author:* *lukman.laliyo@ung.ac.id*
How to cite: L.A.R.Laliyo, Y. Katili, A. Lukum, A.L. Kilo, and M. Pikoli, " Rasch PCM Diagnostic Analysis of Critical Thinking Item Responses in Indonesian Atomic Structure Learning," *Jurnal Kimia dan Pendidikan Kimia (JKPK)*, vol. 10, no. 3, pp.130-146, 2026. [Online]. Available: https://doi.org/10.20961/jkpk.v11i1.112310
Received: 2025-12-09 *Accepted: 2026-04-26* *Published: 2026-04-30*

INTRODUCTION

Critical thinking skills (CTS) are core competencies in 21st-century science education, explicitly identified as one of six dimensions of Indonesia's Pancasila Student Profile within the *Merdeka* Curriculum [1]. Yet Indonesian students' CTS attainment remains concerning: the 2022 PISA placed Indonesia at rank 68 of 81 countries in scientific literacy (mean = 383 vs. OECD average = 485) [2]. Diagnostic studies in chemistry education consistently show serious difficulties in applying scientific reasoning to abstract topics, including atomic structure [3][4] with internationally documented evidence that even targeted pedagogical interventions yield adequate comprehension in only approximately 64% of students with learning difficulties [5]. The development of diagnostic instruments

sensitive to students' CTS on such structural topics thus warrants urgent attention [6]. Measuring CTS on atomic structure is particularly challenging due to persistent misconceptions — electrons "orbiting" the nucleus, atoms equated with "smallest cell parts," or misinterpreted nuclear shape [3][7] that are documented across all educational levels, including among teachers. More fundamentally, CTS is not monolithic: interpretation and representational analysis differ cognitively from evaluation, inference, and self-regulation, the latter involving metacognitive reasoning [8]. These challenges demand instruments that capture reasoning quality, not merely answer correctness.

Assessment development has progressively addressed this. Single-tier tests failed to distinguish guessing from valid reasoning; two-tier tests [9] combined content with reasoning and improved misconception detection [10]; three-tier tests added confidence [11]; and four-tier tests separated content and reasoning confidence [12][13]. Recent systematic reviews confirm four-tier formats' superior diagnostic power for detecting stable misconceptions [14]. Concurrently, the Rasch Partial Credit Model (PCM) has emerged as a measurement approach placing person ability and item difficulty on the same continuum for transparent, fair, sample-independent mapping [15][16], with demonstrated feasibility in Indonesian science education for instrument quality analysis, difficulty pattern evaluation, and DIF detection [17][18]. Two gaps remain inadequately addressed. First, most multi-tier instruments on atomic structure map misconceptions rather than specifically measure CTS — meaning response patterns reflecting reasoning quality have not been systematically analyzed as meaningful CTS diagnostic units. Second, most studies use dichotomous or simple scoring, losing information embedded in partial response patterns that Rasch PCM can preserve through psychometrically meaningful partial credits. Systematic reviews further confirm that four-tier research is rarely combined with polytomous Rasch frameworks for CTS estimation on abstract chemistry topics [14].

This study addresses these gaps by developing and validating a complex multiple-choice instrument on atomic structure with a partial credit scoring strategy, whereby response patterns are mapped as ordered categories according to their CTS inference value. The Rasch PCM was applied to: (1) estimate person ability and item/category difficulty parameters; (2) examine item-person fit; (3) evaluate category function through CPC analysis; and (4) visualize person-item maps (Wright Map) for measurement range assessment. The study aimed to: (i) develop a CTS item bank operationalizing chemistry CTS indicators; (ii) test validity, reliability, and category function via Rasch PCM; (iii) analyze CPC and DIF response patterns; and (iv) present a student ability map for targeted intervention. This article contributes to chemistry assessment methodology by offering an analytical framework for complex multiple-choice diagnostic items aligned with CTS constructs, demonstrating Rasch PCM as a fair polytomous measurement engine, and

providing validity evidence relevant to the Indonesian *Merdeka* Curriculum context.

METHODS

1. Research Design

This study employed a non-experimental quantitative approach with a cross-sectional design. This approach was chosen because the primary purpose was to describe and evaluate students' CTS response patterns on atomic structure at a single point in time, without providing any treatment or intervention to respondents [19]. Data were collected simultaneously in a single measurement period during the even semester of the 2024–2025 academic year. This design is appropriate for instrument evaluation studies aimed at simultaneously estimating person ability and item difficulty parameters on the same continuum [16]. Researchers did not manipulate learning processes or materials; consequently, this study draws no causal conclusions.

The research design logic encompasses three complementary analytical orientations: (1) **descriptive**, to map students' CTS ability profiles based on Facione's six dimensions; (2) **psychometric**, to evaluate instrument quality through Rasch PCM modeling; and (3) **comparative**, to detect differential item functioning (DIF) based on gender and grade level. These three orientations directly map onto research objectives (i)–(iv) stated in the Introduction.

2. Participants and Research Context

Respondents in this study were 850 senior high school students from several schools in Gorontalo Province, Indonesia. School selection was conducted purposively based on accessibility and the required grade-level representation, while respondent selection within each school was performed through simple random sampling from available student lists. Accordingly, the sampling strategy employed was cluster-purposive with within-cluster randomization — neither pure convenience sampling nor full random sampling.

All respondents had received instruction on atomic structure concepts under the *Merdeka* Curriculum. Respondents participated voluntarily and provided written informed consent prior to data collection. No respondents received special instructional treatment in connection with this study. Respondent demographic characteristics are presented in Table 1.

Table 1. Respondent Demographics

Student	Code	Number	
		N	%
Class Level			
Class X	X	519	61
Class XI	Y	150	17,6
Class XII	Z	181	21,3
Gender			
Male	M	351	41,3
Female	F	499	58,7

The sample size of 850 is considered adequate for Rasch PCM analysis, given the recommended minimum of 250–500 respondents for parameter estimation stability in polytomous modeling [20]. The distributional imbalance across grade levels (Grade X representing 61.1%) reflects actual school conditions in the region and was considered in interpreting DIF results by grade level.

3. Conceptual Domain of Atomic Structure

Prior to item development, the conceptual domain of atomic structure material relevant to critical thinking skills (CTS) was mapped. This mapping served as a content anchor to ensure that each diagnostic item measured a specific CTS dimension within a curricularly verified conceptual context. The mapped domains covered four core areas of atomic structure: the development of atomic models, including electron transitions, photon energy, and spectral interpretation in the Bohr model; electron configuration and quantum numbers, including writing electron configurations, determining quantum numbers, and locating elements in the periodic table; isotopes and relative atomic mass, including calculations of relative atomic mass (Ar) based on isotope abundance data; and basic quantum mechanics, including orbital concepts, wave function probability density (ψ^2), and electron probability. This domain mapping provided the foundation for developing the CTS construct, as presented in Table 2.

Table 2. Construct Map of Critical Thinking Skills: Atomic Structure Concept

No	CTS Dimension	Operational Indicator	Measured Aspect	Item Code
1	Interpretation	Clarifying the meaning of a statement	The ability to explain the relationship between photon energy and electron transitions in the Bohr atomic model	Int_01
		Clarifying scientific facts	The ability to understand ionization energy patterns in the periodic table	Int_02
2	Analysis	Examining information/facts	The ability to analyze electron configurations, quantum numbers, and the position of elements in the periodic table	Ana_03
		Verifying information/facts	The ability to analyze the energy of photons emitted from various electron transitions between energy levels	Ana_04
3	Evaluation	Verifying the accuracy of statements	The ability to evaluate the relative atomic mass (*Ar*) based on isotope data	Eva_05
4	Inference	Responding to more than one solution	The ability to deduce the properties of elements from their electron configurations	Inf_06
		Drawing logical conclusions	The ability to distinguish between true and false statements about the fundamental structure of the atom	Inf_07
5	Explanation	Identifying accurate information	The ability to provide correct reasoning about atomic properties based on protons and electrons	Exp_08
6	Self-Regulation	Monitoring and correcting understanding	The ability to correct misconceptions about electron probability in wave mechanics theory	Reg_09
		Validating scientific concepts	The ability to validate the concept of orbitals and the wave function (ψ^2) as the probability of electron existence	Reg_10
Note: CTS = Critical Thinking Skills				

4. Instrument

The measurement instrument in this study was a complex multiple-choice test comprising ten diagnostic items. Each item provided five answer options: two correct answers and three distractors logically designed so that students could not guess the correct answers randomly. Distractors

were constructed based on misconceptions commonly documented in the atomic structure literature [3], thereby increasing the diagnostic power of the instrument [17].

Item scoring employed a three-level partial credit system: (a) score 2: student selects both correct answers; (b) score 1: student selects one correct answer; and (c) score 0: student selects no correct answer. This scoring system enables the capture of partial reasoning information unavailable from conventional dichotomous scoring, and is conceptually aligned with the PCM assumption that responses with higher credits reflect qualitatively superior ability levels [15]. An example of item Int_01, measuring the interpretation ability of the photon energy–electron transition relationship in the Bohr model, is presented in Table 3.

Table 3. Example of Int_01 item

Component	Description
Item code	Int_01
CTS dimension	Interpretation
Conceptual domain	Development of the atomic model
Sub-concept	Photon energy and electron transitions in the Bohr atomic model
Item prompt	Select the correct statements regarding photon energy and electron transitions in the Bohr atomic model.
Option A	The energy of a photon emitted during a transition from a lower energy level to a higher one is always greater. [Incorrect]
Option B	The energy of a photon emitted during a transition from a higher energy level to a lower one is equal to the energy difference between the two levels. [Correct]
Option C	The energy of a photon associated with the transition from n = 1 to n = 3 is smaller than that associated with the transition from n = 3 to n = 1. [Incorrect]
Option D	An electron transition to a higher energy level always requires the absorption of a photon. [Correct]
Option E	Photon energy is not affected by the energy difference between levels. [Incorrect]
Scoring criteria	Selecting both B and D = 2 points; selecting either B only or D only = 1 point; selecting any other response pattern = 0 points.

5. Instrument Validation and Quality

Instrument validation was conducted in two stages: expert content validity and student readability testing. Stage 1 — Expert Validation. Three expert validators with expertise in chemistry education and assessment instrument development were asked to evaluate: (a) item alignment with CTS indicators and dimensions; (b) chemistry content accuracy; (c) language clarity and readability; and (d) distractor quality in detecting misconceptions. Each aspect was rated using a four-point scale (*not appropriate* to *very appropriate*). Inter-validator agreement was analyzed using Fleiss κ, yielding κ = 0.97 (p < 0.0001), classified as *almost perfect agreement* according to Landis and Koch's criteria [21]. All items were declared content-valid and comprehensible, with minor editorial revisions prior to use.

Stage 2 — Readability Test. Prior to main data collection, the instrument was piloted with 30 students outside the main sample to obtain feedback on readability and item comprehension. Results confirmed that all items were well understood, with no students reporting ambiguous instructions or unclear answer choices.

6. Data Collection Procedure

Data collection was conducted during the even semester of the 2024–2025 academic year through several systematic stages. Permission was obtained from school principals, and institutional ethical clearance was secured prior to data collection. Students and parents or guardians then received written explanations regarding the study's purpose, voluntary participation, data confidentiality, and the right to withdraw without any consequence. Written consent was obtained before students completed the instrument. The instrument was administered in written form in each classroom during a single session of approximately 45 minutes under the supervision of the researcher or a briefed classroom teacher. Testing conditions were standardized across schools to minimize situational effects. After the test administration, students' responses were coded using the partial credit scoring system of 0, 1, and 2, and the coded data were entered into a data matrix for further analysis.

7. Ethical Statement

This study was conducted in compliance with ethical principles governing research involving human participants. All participants participated voluntarily based on signed informed consent. Individual student identities and data were kept confidential and used exclusively for aggregate analysis purposes. No procedures posed physical or psychological risks to respondents. The study received approval from the relevant schools and the researcher's institution in accordance with applicable procedures.

8. Data Analysis

Data were analyzed using WINSTEPS software version 4.5.5 [20], the summary statistics are presented in Table 4. The Rasch model was selected for its ability to place person ability and item difficulty on the same latent continuum, enabling interval-based measurement independent of any particular sample [16].

Table 4. Summary Statistic

Psychometric	Person	Item
N	850	10
Mean	.02	.00
SD (Standar Deviasi)	1.03	.69
SE (Standar Error)	.59	.06
Reliability	.61	.99
Separation	1.26	10.63
Infit – MNSQ	.99	.99
Outfit – MNSQ	1.00	1.00
Infit – ZSTD	-.05	-.18
Outfit – ZSTD	-.04	-.02
Cronbach Alpha (KR-20)		.65
Unidimensionality:		
Total Raw Variance in Observations	= 14.5668 (100%).	
Raw Variance Explained by Measures	= 4.5668 (31.4%).	
Raw Unexplained Variance (residual)	= 10.0000 (68.6%).	
Unexplained Variance in 1st Contrast	= 1.5399 (10.6%).	
Unexplained Variance in 2nd–5th Contrast	= 7.4%–9.0%.	
Source: Compiled from WINSTEPS 4.5.5 results		

Analysis was conducted in stages as described below, with explicit mapping to research objectives. Stage 1 — Unidimensionality Test (addressing Objective ii): Evaluated through Principal Component Analysis of Residuals (PCAR). Criteria: (a) variance explained by measures ≥ 20%; and (b) unexplained variance in the first contrast < 15% [22].

Stage 2 — Reliability and Separation (addressing Objective ii): Evaluated at two levels: (a) Rasch-based person and item reliability; and (b) internal consistency via Cronbach's Alpha (KR-20). Separation values determined the instrument's ability to differentiate ability groups. Stage 3 — Item Fit (addressing Objective ii): Evaluated using MNSQ Infit and Outfit values and ZSTD statistics. Items were declared fit when MNSQ ranged within 0.7–1.3 and ZSTD within –2.0 to +2.0 [22][23].

Stage 4 — Category Function via CPC (addressing Objective iii): Category Probability Curve (CPC) was used to verify that all three response categories (0, 1, 2) functioned sequentially and consistently, with each category having a clear dominance region within a specific ability range [20]. Stage 5 — Wright Map (addressing Objective iv): The Wright Map visualized the relative positions of student abilities and item difficulties on the same logit scale, enabling pedagogical interpretation of which CTS dimensions students had mastered and which remained challenging.

Stage 6 — DIF Analysis (addressing Objective iii): Differential Item Functioning (DIF) was tested to detect whether score differences across demographic groups (gender and grade level) reflected genuine ability differences or were artifacts of items functioning differently across groups. DIF was analyzed on the same Rasch logit scale to ensure fair and measurable cross-group comparison [24].

RESULT AND DISCUSSION

1. Rasch Instrument Properties and Psychometric Quality

Results of the Rasch PCM analysis using WINSTEPS 4.5.5 are presented in Table 4. Overall, the instrument demonstrated adequate fit with the Rasch model. MNSQ Infit and Outfit values at both the person level (0.99–1.00) and item level (0.99–1.00) fell within the ideal range of 0.5–1.5, while ZSTD values approached zero (–0.18 to –0.04), indicating no systematic distortion in overall response patterns [23].

Individual item fit evaluation employed the following criteria: (a) Outfit MNSQ within the range 0.7–1.3; values above 1.3 indicate excessive random variation, while values below 0.7 indicate an item that is overly predictable; (b) ZSTD within –2.0 to +2.0; and (c) positive point-measure correlation (PTMEA) as evidence of item consistency with the measured construct [22][23]. Results are presented in Table 5.

Nine of the ten items met all fit criteria and were declared fit with the Rasch PCM. All Outfit MNSQ values ranged within 0.91–1.08, well within the 0.7–1.3 tolerance. One item requiring attention was Reg_09, with MNSQ = 0.91 (within tolerance) but ZSTD = –2.19, slightly exceeding the negative threshold of –2.0. This negative ZSTD indicates an overly predictable response pattern, suggesting that

Reg_09 — measuring the ability to correct electron probability misconceptions — possesses a very dominant correct-answer attractor, driving more uniform responses than the model anticipated. The PTMEA for Reg_09 (0.33) was the lowest among all items, indicating lower discriminatory power, yet remained positive and consistent with the construct direction. All PTMEA values (0.33–0.64) were positive, confirming that all ten items consistently measure the same CTS construct.

These fit results demonstrate adequate psychometric quality across multiple indicators. Item reliability reached a very high level (0.99; separation = 10.63), indicating that the item difficulty hierarchy is highly stable and capable of differentiating more than ten distinct ability groups on the logit scale — a level of psychometric richness rarely achieved with ten-item instruments [16]. The full satisfaction of Rasch fit criteria by all ten items confirms that no item produced response patterns deviating significantly from model expectations, a prerequisite for valid person ability estimation [23]. This contrasts favorably with earlier multi-tier diagnostic studies that reported misfitting items attributable to poorly constructed distractors or item-respondent mismatches [13][17]. The case of Reg_09 warrants more nuanced interpretation. The marginal ZSTD = −2.19 does not constitute traditional misfit; rather, it signals an overly uniform response pattern consistent with a dominant correct-answer attractor. Distractor strengthening is recommended in future instrument revisions to improve its discriminatory power between moderate- and high-ability students.

Table 5. Item Statistics: Misfit Order

Code Item	Measure	SE	Outfit MNSQ	Outfit ZSTD	PTMEA Corr.	Status
Int_02	–0,45	0,06	1,08	1,84	0,64	Fit
Int_01	–0,91	0,06	1,07	1,00	0,61	Fit
Ana_04	–0,17	0,06	1,07	1,58	0,47	Fit
Eva_05	–0,02	0,06	1,06	1,29	0,51	Fit
Inf_07	–0,13	0,06	1,02	0,58	0,44	Fit
Inf_06	0,00	0,06	0,97	–0,77	0,58	Fit
Exp_08	0,02	0,06	0,96	–0,78	0,37	Fit
Ana_03	–0,81	0,06	0,94	–1,32	0,50	Fit
Reg_10	1,32	0,07	0,94	–1,46	0,35	Fit
Reg_09	1,15	0,07	0,91	–2,19	0,33	Requires Attention

Source: Compiled from WINSTEPS 4.5.5 results

Person reliability was moderate (0.61; separation = 1.26; KR-20 = 0.65). This moderate profile must be understood in context and should not be interpreted as a standalone instrument weakness. Three contextual factors explain it: the instrument comprised only ten items; the sample spanned three grade levels with substantially heterogeneous learning experiences; and the Wright Map identifies gaps at very low (< –2 logit) and very high (> +2 logit) ability levels, and in the +0.5 to +1.0 logit range (between Exp_08 and Reg_09), indicating suboptimal continuum coverage. Targeted item additions

at these gaps in future development would directly improve person reliability

2. Unidimensionality

PCAR analysis showed that variance explained by measures was 31.4% exceeding the recommended 20% minimum for educational studies [22]. Unexplained variance in the first contrast was 10.6% and in the second through fifth contrasts ranged from 7.4% to 9.0%. These findings warrant cautious and balanced interpretation. While the 31.4% figure provides initial support for the unidimensionality assumption, the first contrast value of 10.6% cannot be dismissed, as it slightly exceeds the 10% threshold often used to signal a possible secondary dimension [20]. This is substantively understandable: the instrument spans six CTS dimensions forming a theoretical hierarchy from concrete interpretation to metacognitive self-regulation [8].

Facione's six CTS dimensions form a theoretical hierarchy in which interpretation and analysis rely primarily on representational processing, while self-regulation involves a qualitatively distinct metacognitive layer. The 10.6% residual likely reflects genuine cognitive heterogeneity across dimensions rather than measurement noise — these dimensions correlate within a single CTS construct yet each carries unique cognitive demands, contributing to residuals that are not entirely random. Accordingly, unidimensionality is tentatively supported — the instrument is sufficiently valid for measuring one primary CTS construct on atomic structure for practical measurement purposes — while acknowledging that stronger validation with more diverse samples, expanded content domains, homogeneous single-grade samples, and multidimensional confirmatory analysis is needed for stronger confirmation [6].

3. Category Function: Validity of the Partial Credit System via CPC

CPC analysis verified that all three response categories (0, 1, 2) functioned sequentially with distinct dominance regions. Figure 1 presents CPCs for two contrasting items: Int_01 (easiest; measure = –0.91) and Reg_10 (hardest; measure = +1.32) For Int_01, the curves followed an ideal sequential pattern: category 0 dominated at low logits (below –1.5), category 1 had a clear dominance region at moderate logits (–1.5 to 0), and category 2 began to dominate as logit approached and exceeded 0. This confirms that even average-ability students have a high probability of answering completely correctly, consistent with Int_01's position as the easiest item.

For Reg_10, category 2 only dominated at logit values above +1.5 to +2.0, meaning only students with ability well above average could answer completely correctly. Most students stopped at category 1 (partial answer). Notably, Reg_10's MNSQ and ZSTD values (0.94 and –1.46, respectively) suggest that this partial response pattern is highly consistent and predictable — students below the ability threshold uniformly produced partial responses rather than random ones.

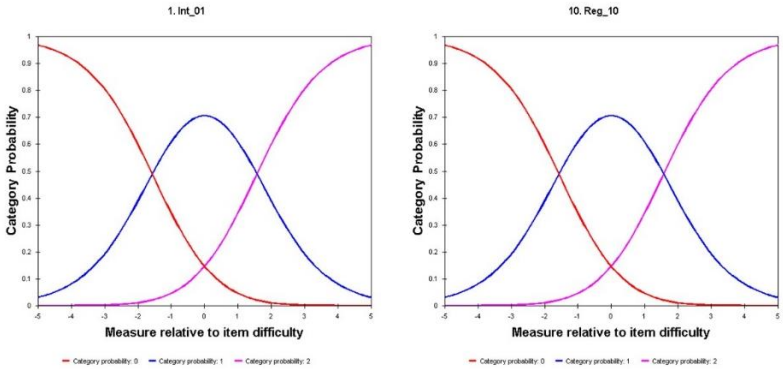


Figure 1. Category Probability Curves for Items Int_01 and Reg_10

Overall, CPC analysis confirmed that all three response categories functioned sequentially across all ten items, without any category collapsing or unreasonably overlapping [27]. This validates the partial credit scoring system (0–1–2) as psychometrically meaningful: each score level captures a qualitatively distinct level of CTS reasoning that dichotomous scoring cannot differentiate [18][28]. The CPC contrast between Int_01 and Reg_10 visually demonstrates the instrument's full diagnostic spectrum, from items accessible to average-ability students to those requiring well-above-average cognitive performance [29].

4. Wright Map: Person Ability, Item Difficulty Distribution, and CTS Hierarchy

Figure 2 presents the Wright Map visualizing the simultaneous distribution of student abilities and item difficulties on the same logit scale, enabling direct pedagogical interpretation of instrument-respondent alignment. Most students were concentrated in the logit range of 0 to +1, placing them at an intermediate ability level. The item distribution on the Wright Map confirms the hierarchical progression of CTS dimensions, from easiest to most difficult based on measure values from Table 5.

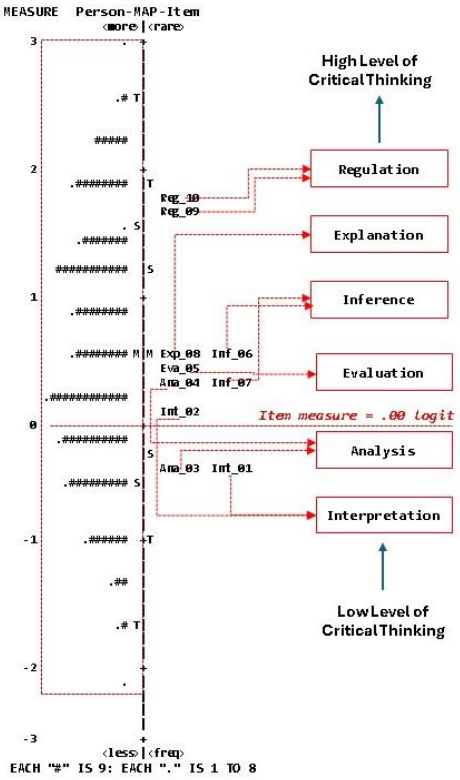


Figure 2. Wright Map Person–Item

This empirical confirmation of a hierarchical CTS progression across the logit scale — converging with Facione's theoretical framework [8] — provides the first Rasch PCM-based evidence of this hierarchy on atomic structure in Indonesian chemistry education. Low Difficulty — Interpretation and

Analysis (logit –0.91 to –0.17): Items Int_01 (–0.91), Ana_03 (–0.81), Int_02 (–0.45), and Ana_04 (–0.17) occupied the lower end of the difficulty scale. This is consistent with their demand for concrete-level representation processing — interpreting photon energy and electron transitions, and analyzing electron configurations. These items reflect intensively practiced curriculum content at concrete representational levels, explaining their relative accessibility — paralleling Chandrasegaran et al. [10], who found descriptive-interpretive abilities precede evaluative abilities developmentally.

Moderate Difficulty — Evaluation, Inference, and Explanation (logit –0.13 to +0.02): Items Inf_07 (–0.13), Eva_05 (–0.02), Inf_06 (0.00), and Exp_08 (+0.02) clustered around the mean student ability (logit = 0.02), constituting the most diagnostically sensitive zone. Students below average still struggle with verifying isotope statement accuracy and drawing conclusions from electron configurations, while above-average students begin constructing systematic scientific arguments. This range is the most informative for instructional intervention.

High Difficulty — Self-Regulation (logit +1.15 to +1.32): Items Reg_09 (+1.15) and Reg_10 (+1.32) were positioned far above mean student ability, representing the cognitive peak. This confirms that metacognitive ability — correcting electron probability misconceptions and validating wave function (ψ^2) concepts in quantum mechanics — requires the highest CTS level. The +1.13 logit gap between Exp_08 (+0.02) and Reg_09 (+1.15) quantifies a genuine cognitive discontinuity: validating orbital concepts and the wave function (ψ^2) demands metacognitive monitoring and self-correction beyond content mastery — abilities requiring explicitly reflective learning rather than procedural drill [12]. This gap prescribes targeted scaffolding through self-explanation, error analysis, and macro–submicro–symbolic multiple representations [25][26].

The Wright Map also identifies two important gaps: (1) no items accommodate very low ability (< –2 logit); and (2) no items exist in the +0.5 to +1.0 logit range between Exp_08 and Reg_09, creating a gap that warrants future instrument development.

5. Differential Item Functioning (DIF): Curriculum Exposure, Not Instrument Bias

DIF testing ensured that score differences across demographic groups genuinely reflected ability differences rather than item artifacts [24]. DIF was analyzed by: gender (male vs. female), and grade level (Grades X, XI, XII). Figure 3 presents the DIF plots for both variables.

DIF by Gender (Figure 3): For items at low to moderate difficulty — Int_01, Ana_03, Int_02, Ana_04, Eva_05, Inf_06, Inf_07, and Exp_08 (measure range –0.91 to +0.02) — difficulty patterns between male and female students were approximately equivalent. No substantial DIF was found, indicating comparable opportunities for both gender groups on these items.

More visible DIF emerged on Reg_09 and Reg_10, where female students in this sample tended to experience relatively greater difficulty. This finding must be

interpreted cautiously: the observed differences cannot be directly interpreted as evidence of gender-based instrument bias, as the instrument did not independently measure variables that could explain these differences — such as cognitive style, exposure to spatial-quantitative exercises, or frequency of symbolic representation-based assignments. Group-based cognitive attributions — such as males' spatial superiority or females' reflective capacity — are methodologically unsupportable, as neither variable was independently measured. This finding is more appropriately interpreted as an early signal that female students in this sample may have had more limited opportunities to practice the mathematical-spatial representations required by self-regulation items [30][31]. Equal representational practice across gender groups should therefore be ensured [14]. Further investigation with designs explicitly controlling for learning exposure variables is warranted.

DIF by Grade Level (Figure 3): Grade-level DIF revealed patterns strongly influenced by curriculum sequence and depth of concept exposure. Items Int_01 and Ana_03 were relatively easier for Grade XI students, likely because they had just completed intensive instruction on these topics before moving to chemical bonding and stoichiometry. Items involving deeper quantum aspects (Ana_04, Inf_06) appeared more challenging for Grade XI students, who had not yet fully re-integrated quantum concepts.

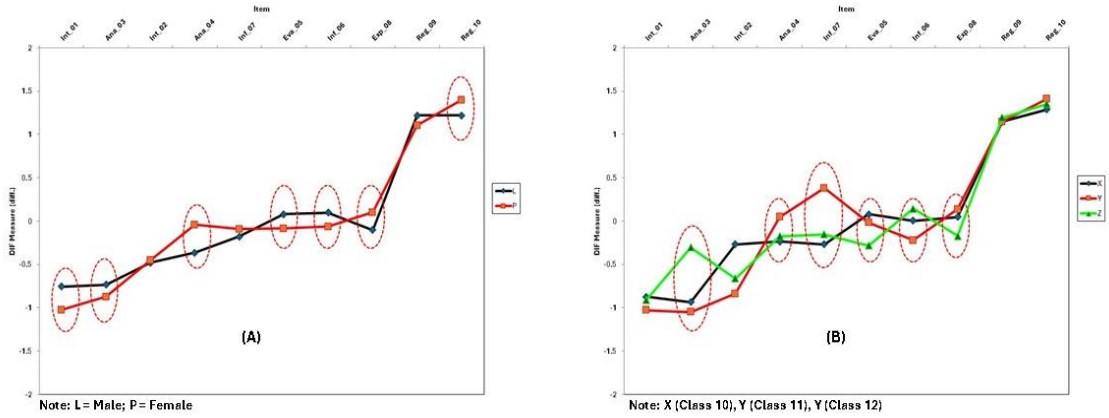


Figure 3. DIF Plots by Gender and Grade Level

Grade XII students performed better on Exp_08, which required scientific justification consistent with their more mature experience in scientific discussion and presentation. Items Reg_09 and Reg_10 posed the greatest challenge for all grade groups, particularly Grade XI, as self-regulation in the quantum mechanics context demands cumulative conceptual maturity that Grade XII students are better positioned to meet.

Overall, DIF findings confirm two important points: first, the instrument is sensitive to pedagogically relevant curriculum exposure differences; second, cross-group performance differences more

likely reflect variations in learning experience than inherent instrument bias. Grade-level DIF, in particular, yields positive content validity evidence: difficulty patterns shifted systematically with curriculum sequencing. This curricular sensitivity constitutes a form of consequential validity rarely documented in chemistry diagnostic instrument studies [32][33], and reinforces the argument that CTS score interpretation must always consider students' curricular context.

6. Comparison with Previous Studies

This study uniquely integrates three rarely combined methodological elements: complex multiple-choice format with explicit partial credit scoring, Rasch PCM modeling preserving response gradation, and logit-based DIF enabling fair cross-group comparison. While two-tier studies [9][10] demonstrated superior misconception detection, their dichotomous scoring sacrificed partial response information. Four-tier studies [12][13] added confidence dimensions but seldom employed polytomous modeling. This study shows that partial credit scoring alone yields meaningful diagnostic gradations at lower respondent cognitive load — without explicit confidence tiers. Compared to prior Indonesian Rasch studies [17][18] focused on simpler conceptual domains, this study extends PCM application to the complex CTS domain with additional CPC and DIF validation evidence.

7. Limitations

Five limitations bound the generalizability of these findings. First, a single-province convenience sample (Gorontalo) constrains external validity. Second, ten items produce suboptimal continuum coverage, particularly in identified logit gaps. Third, cross-sectional design precludes mapping CTS developmental trajectories. Fourth, criterion validity evidence — such as correlation with performance-based CTS assessments — is absent. Fifth, DIF interpretation cannot fully disentangle CTS ability from chemistry content mastery, as both are simultaneously measured.

8. Implications

Theoretically, these findings support reconceptualizing chemistry CTS as a layered competency hierarchy in which representational management (interpretation and analysis) underlies evaluation and inference, while metacognitive self-regulation demands orchestration of all preceding layers. The empirically quantified logit gap (+1.13) between explanation and self-regulation confirms that metacognitive CTS does not develop linearly but requires qualitatively distinct pedagogical conditions [8]. Logit-based DIF analysis also emerges as a curricular context validation tool, extending its function beyond conventional bias detection.

Practically, teachers can use the Wright Map CTS profile to precision-target instruction across three zones: students at negative logits require reinforcement of concrete macro–submicro–symbolic representations; those at moderate logits benefit from structured scientific argumentation; and the +1.13 logit gap signals an urgent need for explicit metacognitive scaffolding — through self-explanation and error analysis — between

explanation and self-regulation. For instrument developers, the highest priority is adding items within the +0.5 to +1.0 logit gap and at scale extremes (< −1.5 and > +2.0 logit) to improve continuum coverage and person reliability.

CONCLUSION

This study developed and validated a Rasch PCM-based complex multiple-choice diagnostic instrument to measure critical thinking skills (CTS) in atomic structure concepts among 850 senior high school students in Gorontalo Province, Indonesia. Four principal conclusions follow.

First, ten items successfully operationalized Facione's six CTS dimensions — interpretation, analysis, evaluation, inference, explanation, and self-regulation — in a hierarchical and structured manner. Item positions on the logit scale ranged from −0.91 (interpretation) to +1.32 (self-regulation), confirming a progressive difficulty structure aligned with Facione's theoretical framework.

Second, psychometric evaluation confirmed adequate instrument quality. Item reliability was very high (0.99; separation = 10.63); all ten items met fit criteria (Outfit MNSQ = 0.91–1.08). Reg_09 (ZSTD = −2.19) requires distractor revision to improve discriminatory power. Person reliability was moderate (0.61; KR-20 = 0.65), reflecting sample heterogeneity across three grade levels. Unidimensionality was tentatively supported (variance explained = 31.4%), with the first contrast of 10.6% indicating a possible secondary dimension consistent with the hierarchical nature of the six CTS components.

Third, response pattern analyses via CPC and DIF yielded complementary validity evidence. CPC confirmed that all three response categories (0, 1, 2) functioned sequentially across all items, validating the partial credit system as psychometrically meaningful. DIF revealed that cross-group performance differences reflect curriculum exposure variations rather than instrument bias, establishing the instrument's sensitivity to genuine learning effects as a form of consequential validity.

Fourth, the Wright Map identified a logit gap of +1.13 between Exp_08 (+0.02) and Reg_09 (+1.15), confirming a genuine cognitive discontinuity between explanation and metacognitive self-regulation in quantum mechanics. This gap provides a precise target for pedagogical scaffolding intervention.

Collectively, these findings establish Rasch PCM as an appropriate framework for evaluating partial credit-based CTS diagnostic instruments in chemistry education. The instrument captures reasoning gradations through validated sequential response categories, providing richer diagnostic information than dichotomous instruments — a claim grounded in internal psychometric evidence and subject to confirmation through future comparative studies.

Primary limitations include a single-province sample, a restricted ten-item pool, cross-sectional design, and the absence of external criterion validity. Future research priorities are: (1) expanding the item bank,

particularly at the +0.5 to +1.0 logit gap and scale extremes; (2) cross-regional validation for measurement invariance; (3) longitudinal designs to map CTS developmental trajectories from Grade X to XII; (4) mixed-method investigation — cognitive interviews and written explanation analysis — to examine the self-regulation bottleneck; and (5) criterion validity development through correlation with authentic performance-based CTS assessments.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the financial support provided by the Directorate of Research and Community Service (Direktorat Riset dan Pengabdian Masyarakat—DRPM), Ministry of Higher Education, Science, and Technology of the Republic of Indonesia. This research was funded under the Rector Decree of Universitas Negeri Gorontalo No. 987/UN47/HK.02/2025, the Main Contract No. 082/C3/DT.05.00/PL/2025, and the Derivative Contract No. 689/UN47.D1/PT.01/2025.

REFERENCES

[1] Kemdikbudristek, "Keputusan Kepala BSKAP Nomor 033/H/KR/2022 tentang Capaian Pembelajaran pada Kurikulum Merdeka," Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, Jakarta, 2022.

[2] OECD, PISA 2022 Results (Volume I): The State of Learning and Equity in Education, OECD Publishing, Paris, 2023. https://doi.org/10.1787/53f23881-en

[3] A. K. Griffiths and K. R. Preston, "Grade-12 students' misconceptions relating to fundamental characteristics of atoms and molecules," Journal of Research in Science Teaching, vol. 29, no. 6, pp. 611–628, 1992. https://doi.org/10.1002/tea.3660290609

[4] K. S. Taber, "Alternative conceptions in chemistry: Prevention, diagnosis and cure?" School Science Review, vol. 83, no. 304, pp. 92–101, 2002.

[5] G. López-Íñiguez, A. Pontes-Pedrajas, and R. E. Valle-Flórez, "'Atomizados': An educational game for learning atomic structure. A case study with grade-9 students with difficulties learning chemistry," Journal of Chemical Education, vol. 100, no. 8, pp. 3114–3123, 2023. https://doi.org/10.1021/acs.jchemed.2c00614

[6] L. Laliyo, R. Utina, R. Husain, M. K. Umar, M. R. Katili, and C. Panigoro, "Evaluating students' ability in constructing scientific explanations on chemical phenomena," Eurasia Journal of Mathematics, Science and Technology Education, vol. 19, no. 9, em2326, 2023. https://doi.org/10.29333/ejmste/13524

[7] M. B. Nakhleh, "Why some students don't learn chemistry: Chemical misconceptions," Journal of Chemical Education, vol. 69, no. 3, pp. 191–196, 1992. https://doi.org/10.1021/ed069p191

[8] P. A. Facione, "Critical thinking: What it is and why it counts," Insight Assessment, pp. 1–28, 2011.

[9] D. F. Treagust, "Development and use of diagnostic tests to evaluate students' misconceptions in science," International Journal of Science Education, vol. 10, no. 2, pp. 159–169, 1988. https://doi.org/10.1080/0950069880100204

[10] A. L. Chandrasegaran, D. F. Treagust, and M. Mocerino, "The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school

students' ability to describe and explain chemical reactions using multiple levels of representation," Chemistry Education Research and Practice, vol. 8, no. 3, pp. 293–307, 2007. https://doi.org/10.1039/B7RP90006F

[11] H. O. Arslan, C. Cigdemoglu, and C. Moseley, "A three-tier diagnostic test to assess pre-service teachers' misconceptions about global warming, greenhouse effect, ozone layer depletion, and acid rain," International Journal of Science Education, vol. 34, no. 11, pp. 1667–1686, 2012. https://doi.org/10.1080/09500693.2012.680618

[12] I. S. Caleon and R. Subramaniam, "Do students know what they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions," Research in Science Education, vol. 40, no. 3, pp. 313–337, 2010. https://doi.org/10.1007/s11165-009-9122-4

[13] Habiddin and E. M. Page, "Development and validation of a four-tier diagnostic instrument for chemical kinetics (FTDICK)," Indonesian Journal of Chemistry, vol. 19, no. 3, pp. 720–736, 2019. https://doi.org/10.22146/ijc.39218

[14] N. Ö. Çelikkanlı and H. Ş. Kızılcık, "A review of studies about four-tier diagnostic tests in physics education," Journal of Turkish Science Education, vol. 19, no. 4, pp. 1291–1311, 2022. https://doi.org/10.36681/tused.2022.175

[15] G. N. Masters, "A Rasch model for partial credit scoring," Psychometrika, vol. 47, no. 2, pp. 149–174, 1982. https://doi.org/10.1007/BF02296272

[16] T. G. Bond, Z. Yan, and M. Heene, Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 4th ed., Routledge, 2021. https://doi.org/10.4324/9780429030499

[17] L. A. R. Laliyo, J. S. Tangio, B. Sumintono, M. Jahja, and C. Panigoro, "Analytic approach of response pattern of diagnostic test items in evaluating students' conceptual understanding of characteristics of particle of matter," Journal of Baltic Science Education, vol. 19, no. 5, pp. 824–841, 2020. https://doi.org/10.33225/jbse/20.19.824

[18] L. Laliyo, S. Hamdi, M. Pikoli, R. Abdullah, and C. Panigoro, "Implementation of four-tier multiple-choice instruments based on the partial credit model in evaluating students' learning progress," European Journal of Educational Research, vol. 10, no. 2, pp. 825–840, 2021. https://doi.org/10.12973/EU-JER.10.2.825

[19] M. S. Setia, "Methodology series module 3: Cross-sectional studies," Indian Journal of Dermatology, vol. 61, no. 3, pp. 261–264, 2016. https://doi.org/10.4103/0019-5154.182410

[20] J. M. Linacre, A User's Guide to WINSTEPS/MINISTEP Rasch-Model Computer Programs: Program Manual 5.10.2, Winsteps.com, 2025. https://www.winsteps.com

[21] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977. https://doi.org/10.2307/2529310

[22] B. Sumintono, "Rasch model measurement in social science research," MATEC Web of Conferences, vol. 200, art. 01010, 2018. https://doi.org/10.1051/matecconf/201820001010

[23] W. J. Boone, "Rasch analysis for instrument development: Why, when, and how?" CBE—Life Sciences Education, vol. 15, no. 4, art. rm4, 2016. https://doi.org/10.1187/cbe.16-04-0148

[24] Q. Duan and Y. Cheng, "Detecting Differential Item Functioning Using Response Time," Educational and Psychological Measurement, vol. 85, no. 2, pp. 291–312, 2025. https://doi.org/10.1177/00131644241280400

[25] K. Vo, M. Sarkar, P. J. White, and E. Yuriev, "Development of problem-solving skills supported by metacognitive scaffolding: insights from students' written work," Chemistry Education Research and Practice, vol. 25, pp. 1197–1209, 2024. https://doi.org/10.1039/D3RP00284E

[26] J. Laohapornchaiphan and P. Chenprakhon, "A review of research on learning activities addressing the submicroscopic level in chemistry," Journal of Chemical Education, vol. 101, no. 11, pp. 4552–4565, 2024. https://doi.org/10.1021/acs.jchemed.4c00156

[27] L. Tesio, A. Caronni, D. Kumbhare, and S. Scarano, "Interpreting results from Rasch analysis 1. The 'most likely' measures coming from the model," Disability and Rehabilitation, vol. 46, no. 3, pp. 591–603, 2024. https://doi.org/10.1080/09638288.2023.2169771

[28] S. A. Wind, "Detecting rating scale malfunctioning with the partial credit model and generalized partial credit model," Educational and Psychological Measurement, vol. 83, no. 5, pp. 953–983, 2023. https://doi.org/10.1177/00131644221116292

[29] T. A. May, K. L. K. Koskey, J. D. Bostic, G. E. Stone, L. M. Kruse, and G. Matney, "Examining how using dichotomous and partial credit scoring models influence sixth-grade mathematical problem-solving assessment outcomes," School Science and Mathematics, vol. 123, no. 2, pp. 54–67, 2023. https://doi.org/10.1111/ssm.12570

[30] T. Liu and K. Ercikan, "Investigating differential item functioning across interaction variables in listening comprehension assessment," System, vol. 121, art. 103253, 2024. https://doi.org/10.1016/j.system.2024.103253

[31] K. A. Bartlett and J. D. Camba, "Gender differences in spatial ability: A critical review," Educational Psychology Review, vol. 35, art. 8, 2023. https://doi.org/10.1007/s10648-023-09728-2

[32] Y. Dong, D. Dumas, D. H. Clements, C. A. Day-Hess, and J. Sarama, "Evaluating the consequential validity of the research-based early mathematics assessment," Journal of Psychoeducational Assessment, vol. 41, no. 5, 2023. https://doi.org/10.1177/07342829231165812

[33] D. Dumas, Y. Dong, and D. McNeish, "How fair is my test? A ratio coefficient to help represent consequential validity," European Journal of Psychological Assessment, vol. 39, no. 6, pp. 416–423, 2023. https://doi.org/10.1027/1015-5759/a000724