Biosfer: Jurnal Pendidikan Biologi

Journal homepage: http://journal.unj.ac.id/unj/index.php/biosfer

# Psychometric validation of rasch model for critical reasoning on virus material in high school biology

**Nani Aprilia[1]*, Hani Irawati[1], Inrayati Atjo[2]**

[1] Biology Education, Faculty of Teacher Training and Education, Ahmad Dahlan University, Indonesia
[2] Student Methodology and Technology of Professional Education, Faculty of Pedagogy, Far Eastern Federal University, Russia.

*Corresponding author: nani.aprilia@pbio.uad.ac.id

**ARTICLE INFO**

**Article history**
Received: 08 May 2026
Revised: 27 May 2026
Accepted: 09 July 2026

**Keywords**:
Deep learning
Higher order thinking
Partial credit model
Viruses

**ABSTRACT**

Critical reasoning assessment is crucial in supporting deep learning at the high school level, especially in viruses. However, valid and reliable Rasch-based assessment instruments are still very limited. This study aims to validate a critical reasoning assessment instrument based on the Partial Credit Model (PCM) Rasch on viruses in high school biology. The research used a research and development approach with the Plomp model involving 105 Class X students at state high schools in Banguntapan district through purposive sampling.The instrument consisted of 15 polytomous items (scale 0–3) covering four aspects of critical reasoning. Data analysis was conducted using Winsteps software with a focus on item fit (Infit/Outfit MNSQ), difficulty parameters (eta & beta), threshold, separation reliability, and Wright Map. The analysis results showed a separation reliability of 0.7677 (fairly good category). Most items had good variation in difficulty levels, although many items showed disordered threshold patterns. Items P13, P14, and P15 were the most difficult items, while P6, P7, and P8 were the easiest. The findings showed that the instrument was able to differentiate student abilities but still needed revision in the response categories. Recommendations on instrument improvements through simplification of response categories and item writing are needed before widespread use. The instrument has the potential to be a formative diagnostic tool that supports an in-depth learning approach in Biology learning.

## INTRODUCTION

Educational evaluation is a dynamic and increasingly complex global issue in the 21st century. Assessment is no longer simply a measure of academic achievement but has become a key catalyst for developing students' higher-order thinking skills (Golden, 2023; Vlachopoulos & Makri, 2024). Amidst massive technological disruption and artificial intelligence, the greatest challenge in education is migrating from traditional memorization-based assessments to holistic assessment models that support deep learning, ensure fair accountability, and minimize negative psychological impacts such as test anxiety (Aristizábal Gómez et al., 2025; Kusurkar et al., 2023; Rodríguez Rincón et al., 2024). The paradigm shift from summative evaluation to feedback-oriented formative evaluation has become a global trend (Ismail et al., 2022; Pat-El et al., 2024). Therefore, the reconstruction of valid, reliable, and adaptive evaluation instruments is absolutely urgent to bridge the gap in educational quality, particularly in understanding global health issues post-pandemic (Kassiavera et al., 2024; Sartono et al., 2025).

Several previous studies have attempted to develop critical reasoning assessment instruments in science and biology learning worldwide using Item Response Theory (IRT), specifically the Rasch model. The Rasch model is widely chosen by international researchers due to its sample-free nature and ability to produce objective, linear measurements equivalent to physical scales (Heim et al., 2023; Hu & Bi, 2025; Micoloi et al., 2025). Research in the United States and China has shown that Rasch-based instruments demonstrate high reliability and strong correlations with students' scientific literacy achievement (Heim et al., 2023; Hu & Bi, 2025). Meanwhile, in Indonesia, the development of Rasch instruments is gaining ground, as demonstrated by Prasetya & Pratama (2023) and Sujatmika et al (2025) on the digestive and circulatory systems. However, most of these studies report similar methodological limitations, namely unequal distribution of item difficulty levels and neglect of students' self-regulation in critical reasoning (Laliyo et al., 2022; Mu'aziyah et al., 2025; Yokhebed et al., 2025).

Although the literature on the Rasch model in science education continues to grow, there is a significant and up-to-date research gap regarding critical reasoning assessments in virology (viruses) at the high school level. The majority of existing instruments still focus on anatomical biology (such as organ systems) and rarely address the domain of viruses, which is crucial in the post-pandemic era to counter medical misinformation (Prasetya & Pratama, 2023; Sujatmika et al., 2025). Furthermore, previous assessments have not comprehensively integrated the four aspects of Paul-Elder and Facione's critical reasoning (asking meaningful questions, clarifying information, distinguishing fact-opinion-inference, and reflective thinking) within a deep learning framework (Fitrianto & Hidayat, 2024; Setiowati et al., 2025). Psychometric validation using the Rasch Partial Credit Model (PCM), which is ideal for measuring gradations of reasoning ability through polytomous scoring, is also almost non-existent in developing countries like Indonesia (Kassiavera et al., 2024).

Therefore, the novelty of this research lies in the development and validation of a Rasch PCM-based critical reasoning instrument that fully integrates the Paul Elder and Facione thinking model into high school viruses. This integration is strategic for accurately detecting and reducing students' biological misconceptions through objective data visualization (Calma & Davies, 2025; Zhao et al., 2026). Furthermore, the research provides a methodological contribution through in-depth Rasch analysis, not only mapping the distribution of student abilities but also evaluating the instrument's effectiveness in supporting in-depth learning, a rarely done approach in virus research (Hu & Bi, 2025; Prasetya & Pratama, 2023). This approach is relevant because it combines the local Indonesian context with global health issues, thus potentially serving as an international reference for the development of future science assessments (Heim et al., 2023; Micoloi et al., 2025).

This study aims to test and validate the psychometric quality of a critical reasoning assessment instrument based on the Rasch Partial Credit Model (PCM) on viruses to support deep learning for high school students. Specifically, this study was designed to answer the following crucial questions: 1. What is the quality of item fit, as represented by the Outfit and Infit Mean Square (MSQ) values, on the critical reasoning assessment instrument on viruses? 2. What is the distribution of the item difficulty (beta) and category threshold (eta) parameters of the instrument developed based on the Rasch model estimation? 3. What is the extent of the separation reliability indices for both the person and item of the instrument?

## METHODS

### Research Design

This study employs a Research and Development (R&D) approach based on the Plomp model, comprising preliminary research, prototyping, and an assessment phase (Akker et al., 2013; Wahyuni et al., 2025). A quantitative method is applied to analyze the instrument's psychometric properties. This approach is selected because the Plomp framework is systematically proven and widely implemented in engineering and science education assessments (Siswanto & Peni, 2023; Sutiarso et al., 2022; Thalhah et al., 2022) . The Plomp model procedure is in Figure 1.

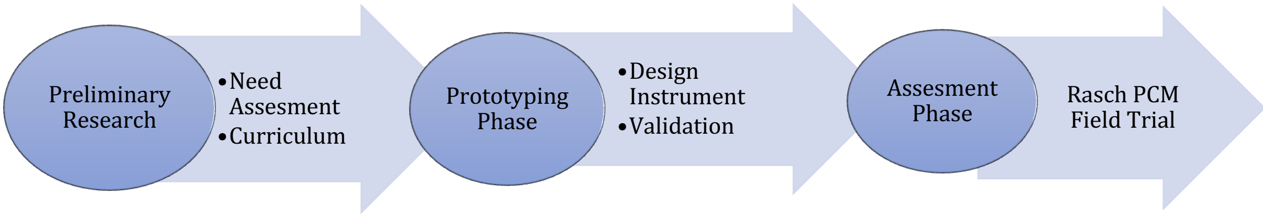


**Figure 1**. Plomp model procedure.

### Research Subjects and Objects

This research was conducted at a state high school in Banguntapan District. The location selection technique used purposive sampling because it represents the suburban-rural area and is a pilot project school representing Banguntapan District. In Table 1, the research subjects were 10th-grade students from a state high school in Banguntapan District, selected purposively, for a total sample of 105 students. The object of this study was students' critical reasoning during in-depth biology learning on viruses (Kassiavera et al., 2024). This sampling technique was relevant because it ensured that participants had homogeneous characteristics and were relevant to students participating in the virus learning, thereby ensuring that the data accurately reflected critical reasoning abilities.

**Table 1**

Research Subjects.

| No | School Name | Class | Σ Population | Male | Female | Σ Sample |
|---|---|---|---|---|---|---|
| 1 | State high school in Banguntapan District (X) | X1-X4 | 144 | 22 | 30 | 52 |
| 2 | State high school in Banguntapan District (Y) | X1-X8 | 288 | 24 | 29 | 53 |
| | | | | | | 105 |

### Instrument

Data collection was conducted through a test (Suwita et al., 2024). This method is relevant because the test directly measures critical reasoning abilities. Table 2 shows the grid of the critical reasoning instrument based on four aspects of critical reasoning (asking meaningful questions, clarifying information, analyzing and evaluating arguments, and reflecting on one's own thinking) (Prasetya & Pratama, 2023). Table 2 presents the instrument, consisting of 15 questions related to the virus material and Bloom's cognitive learning outcomes. The questions were polytomous, with a scale of 0–3 (three-level). The instrument was content validated by experts and analyzed using Aiken V.

**Table 2**

Grid of question instruments based on critical reasoning aspects.

| No | Critical Reasoning Aspects | Material Aspects | Bloom's Level |
|---|---|---|---|
| 1 | Analyzing and Evaluating Arguments | Virus Shape & Structure | C5 |
| 2 | Clarifying Information & Distinguishing Fact-Opinion-Inference | Virus Shape & Structure | C4 |
| 3 | Reflecting and Refining One's Own Thinking | Virus Shape & Structure | C6 |
| 4 | Asking Meaningful Questions | Definition & Characteristics | C4 |
| 5 | Analyzing and Evaluating Arguments | Definition & Characteristics | C5 |
| 6 | Clarifying Information & Distinguishing Fact-Opinion-Inference | Virus Replication | C4 |
| 7 | Analyzing and Evaluating Arguments | Virus Replication | C5 |
| 8 | Reflecting and Refining One's Own Thinking | Virus Replication | C6 |
| 9 | Asking Meaningful Questions | Virus Prevention | C4 |

| No | Critical Reasoning Aspects | Material Aspects | Bloom's Level |
|---|---|---|---|
| 10 | Analyzing and Evaluating Arguments | Virus Prevention | C5 |
| 11 | Reflecting and Refining One's Own Thinking | Virus Prevention | C6 |
| 12 | Clarifying Information & Distinguishing Fact-Opinion-Inference | Virus Utilization | C4 |
| 13 | Analyzing and Evaluating Arguments | Virus Utilization | C5 |
| 14 | Asking Meaningful Questions | Virus Utilization | C4 |
| 15 | Reflecting and Refining One's Own Thinking | Virus Utilization | C6 |

### Procedure

The research procedure was carried out sequentially through three stages of Plomp's model. The first stage, preliminary research, included needs analysis, a literature review, and identification of critical reasoning problems in the virus material (Tavares et al., 2021). The second stage of product development includes the design, creation, and revision of Rasch-based assessment instruments. The third stage of assessment involves field trials, data collection, and evaluation of the instrument's effectiveness (Sujatmika et al., 2025). This procedure is relevant because it ensures that the instrument is developed iteratively and can be replicated in accordance with development research standards.

### Data Analysis Techniques

Quantitative descriptive research data analysis techniques, with Partial Credit Model (PCM) Rasch analysis, including fit (Outfit & Infit MSQ), step difficulty (eta), item easiness (beta), location parameters, category thresholds, and separation reliability using the latest version of Winsteps software (Boone & Staver, 2020; Davis & Boone, 2021). This Rasch analysis is relevant because it can provide objective estimates of student ability and item quality and produce accurate interval measurements for critical reasoning assessments. The acceptance criteria for the Partial Credit Model (PCM) Rasch analysis in this study refer to international standards. The Outfit Mean Square (MSQ) and Infit Mean Square (MSQ) values are accepted within the range of 0.70–1.30, with a maximum tolerance of 0.50–1.50 (Boone & Staver, 2020; Wind, 2023). The step difficulty (eta) and item easiness (beta) parameters are expected to show a minimum variation of 0.5 logit, while the item locations are ideally spread in the range of -3.0 to +3.0 logit (Boone & Staver, 2023). In addition, the category thresholds should increase monotonically with an inter-threshold distance of at least 1.0 logit; the presence of disorganized thresholds is considered non-ideal and requires revision (Kwon et al., 2025; Tesio et al., 2024; Wind, 2023). Person separation reliability is at least 0.70, and item reliability is at least 0.90 (Prasetya & Pratama, 2023; Sujatmika et al., 2025). These criteria serve as the primary reference for assessing the psychometric quality of Rasch-based critical reasoning assessment instruments.

## RESULTS AND DISCUSSION

Development of a Rasch-based critical reasoning assessment for in-depth biology learning of virus material in Plomp model high schools, consisting of three main stages:

1. Preliminary Research Stage

At this stage, a literature review, questionnaire distribution, and in-depth interviews were conducted. The literature review revealed that biology learning in high school is still dominated by evaluation instruments that measure lower-level cognitive abilities, resulting in students' critical reasoning not developing optimally (Aggraini et al., 2024; Yokhebed, Karmadi, & Nastiti, 2025). An analysis of the Independent Curriculum shows that the Learning Outcomes (CP) and Learning Objectives (TP) for the 11th-grade Biology subject in the topic of viruses explicitly require students to develop higher-order thinking skills, namely analyzing the structure and life cycle of viruses, evaluating the impact of virus mutations on health, and reflecting on the relationship between knowledge of viruses and social and environmental issues. However, the implementation of CP and TP has not been supported by adequate assessment instruments, because most evaluation tools still focus on mastery of basic concepts and have not systematically measured the four aspects of critical reasoning (asking meaningful questions, clarifying ideas, analyzing and evaluating, and reflecting on one's own thinking) (Suwita et al., 2024; Usman et al., 2024; Wuisan et al., 2024).

A questionnaire distributed to 35 10th-grade students at State High School in Banguntapan District X and Y found that most students experienced difficulties with critical reasoning on the topic of viruses, with 70–80% stating that the assessment questions focused more on concept mastery than on critical thinking. In-depth interviews with three teachers revealed that they still struggle to develop and implement critical thinking skills, due to various factors, including time constraints. This finding is supported by other research indicating that teachers recognize the importance of critical reasoning but struggle to design systematic instruments due to limited time, resources, and knowledge of modern psychometric methods, such as the Rasch model (Fu & Koh, 2026; Hidayati et al., 2023). Based on these findings, it is concluded that a specific critical reasoning assessment instrument for viruses, analyzed using the Rasch model, is needed and supports in-depth learning.

2. Prototyping Phase

Based on the results of the initial investigation, an instrument was developed consisting of 15 multiple-choice polytomous items with a 0–3 scale (three-tier type). Covering four aspects of critical reasoning (asking meaningful questions, clarifying information and ideas, analyzing and evaluating arguments, and reflecting on one's own thoughts) in the context of virus material. The development process involved expert validation by three biology lecturers, one educational evaluation lecturer, three biology education students, and one 10th-grade high school teacher. The validation results showed an average Aiken V score of 0.87 (valid category). Three items were revised to improve the clarity and suitability of the material. The prototype instrument was revised to achieve expert agreement before being piloted.

3. Assessment Phase

In the assessment phase of the Plomp model, a critical reasoning assessment instrument based on the Partial Credit Model (PCM) Rasch, consisting of 15 polytomous items with a scale of 0–3, was tested on 105 students of SMA Negeri 1 Banguntapan and SMA Negeri 1 Jetis in Bantul Regency. Data analysis was conducted comprehensively to answer the research objective, namely, to determine the psychometric quality of the critical reasoning assessment instrument based on the Partial Credit Model Rasch, using the virus material for senior high school students.

Based on data analysis, information was obtained for each question item with a score range between 0-3. The average item value ranged from 1.34 to 2.13, indicating a moderate response tendency. The item with the highest average was P12 (mean = 2.13), followed by P1 (mean = 2.04), indicating that these items were relatively easier. In contrast, the item with the lowest mean score was P13 (mean = 1.34), followed by P10 (mean = 1.38), indicating that these items were relatively more difficult (Sari & Rakhmawati, 2024). In terms of data distribution, the standard deviation values ranged from 0.97 to 1.19, indicating that the variation in respondents' answers was quite large and relatively even across items. The highest standard deviations were observed for P11 (SD = 1.19) and P13 (SD = 1.18), indicating that respondents' responses to these items were more heterogeneous. Conversely, the lowest standard deviation was observed for P7 (SD = 0.97), indicating that respondents' responses to this item were more homogeneous than those to the other items. (Prasetya & Pratama, 2023; Sujatmika et al., 2025; Tesio et al., 2024; Yokhebed, Karmadi, Silaen, et al., 2025; Yokhebed, Karmadi, & Nastiti, 2025).

The separation reliability is 0.7677, indicating that the instrument has an adequate level of consistency for measuring critical reasoning, although the reliability is still at the lower limit and remains in the good category (Suwita et al., 2024). This value reflects the instrument's ability to meaningfully differentiate students with different levels of critical reasoning, in accordance with the basic principles of Rasch theory, which emphasize invariant measurement and person separation as the main indicators of measurement quality (Alnahdi et al., 2025; Yang et al., 2023). In Rasch analysis, separation reliability indicates the extent to which the instrument consistently distinguishes between respondents with high and low ability. A value of 0.7677 indicates that approximately 76.77% of the observed score variation can be explained by differences in actual ability among respondents, while the remainder is due to measurement error. Therefore, this instrument is suitable for analysis, as it provides relatively stable measurement results (Fitrah et al., 2024; Mallinson et al., 2022). However, the reliability value is not yet very high, so there is still room for improvement in the instrument's quality. These results indicate that although the items are sufficiently capable of differentiating respondents, their separation power is not yet optimal. This condition may be related to several previous findings, such as the presence of unordered thresholds in several items, indicating that the response categories are not yet functioning consistently. Therefore, the instrument can be considered generally quite reliable, but revisions to some items or simplification of the response categories have the potential to improve separation reliability.

In general, Table 3 shows the results of the partial credit model (PCM) analysis, showing variations in the level of difficulty between steps in each item. The eta (Estimate) parameter reflects the level of difficulty in moving from one response category to a higher category. Negative eta values indicate relatively easy steps, while positive eta values indicate more difficult steps. Table 1 shows that the instrument has a fairly good variation in the level of difficulty of steps. The easiest steps are found in items P7.c2 (-1.064), P8.c1 (-0.959), P3.c1 (-0.834), P1.c3 (-0.769), P12.c3 (-0.722), and P6.c2 (-0.721). These very negative eta values indicate that students relatively easily move to higher response categories on these items. This reflects that high school students are more proficient in low to intermediate levels of critical reasoning in the context of virus material. In contrast, the most difficult steps were found in items P13.c3 (1.465), P15.c3 (1.306), P14.c3 (1.045), P10.c3 (1.001), P13.c1 (0.989), and P14.c2 (0.846). These high positive eta values indicate that moving to the highest category (score 3) requires a much higher level of critical reasoning ability. Items P13, P14, and P15 are overall the most difficult group of items in this instrument. Substantively reflects the complexity of the reasoning aspects tested, namely the dimensions of evaluation and reflection on the impact of virus mutations on global health and socio-environmental issues. This aligns with the Learning Outcomes (CP) requirements in the Independent Curriculum, which mandates the development of independent critical thinking and Higher Order Thinking Skills (HOTS) at the metacognitive level (BSKP, 2022). However, the failure of the majority of students to reach the highest score threshold on these items reveals an instructional gap in practice. Although the curriculum demands high levels of reflection, classroom learning rarely systematically trains complex contextual problem-solving due to teachers' time constraints and a lack of references to modern psychometric-based instruments (Retnawati et al., 2016). Substantively, this pattern shows that the instrument can differentiate students with low, medium, and high critical reasoning abilities. Easy items serve as anchors for students with low abilities, while difficult items serve as discriminators for students with high abilities (Kassiavera et al., 2024; Laliyo et al., 2022). These findings support the psychometric quality of the instrument, although the steps on some difficult items (especially P13) still require further evaluation to be more optimal in measuring in-depth critical reasoning on virus material. (Arjana et al., 2025; Micoloi et al., 2025).

Table 3 shows that the easiness (beta) confirms the findings in the difficulty table. The easiest steps have the largest positive betas, namely P7.c2 = 1.064, P8.c1 = 0.959, P3.c1 = 0.834, P1.c3 = 0.769, P12.c3 = 0.722, and P6.c2 = 0.721. This means that the respondents' chances of passing these steps are relatively higher. Conversely, the least easy or most difficult steps appear to have the largest negative betas in absolute terms, namely P13.c3 = -1.465, P15.c3 = -1.306, P14.c3 = -1.045, P10.c3 = -1.001, P13.c1 = -0.989, and P14.c2 = -0.846.

**Table 3**
Results of Eta (Difficulty) and Beta (Easiness) Parameter Analysis.

| Rank | Item Step | Eta (Difficulty) | Beta (Easiness) | Difficulty level | Interpretation |
|---|---|---|---|---|---|
| 1 | P7.c2 | -1.064 | 1.064 | Very Easy | Students move very easily to a higher category |
| 2 | P8.c1 | -0.959 | 0.959 | Very Easy | Students move very easily |
| 3 | P3.c1 | -0.834 | 0.834 | Easy | Students move easily |
| 4 | P1.c3 | -0.769 | 0.769 | Easy | Students move easily |
| 5 | P12.c3 | -0.722 | 0.722 | Easy | Students move easily |
| 6 | P6.c2 | -0.721 | 0.721 | Easy | Students move easily |
| 1 | P13.c3 | 1.465 | -1.465 | Very Difficult | Requires very high critical reasoning skills |
| 2 | P15.c3 | 1.306 | -1.306 | Very Difficult | Requires very high skills |
| 3 | P14.c3 | 1.045 | -1.045 | Difficult | Requires high skills |
| 4 | P10.c3 | 1.001 | -1.001 | Difficult | Requires high skills |
| 5 | P13.c1 | 0.989 | -0.989 | Difficult | Requires high skills |
| 6 | P14.c2 | 0.846 | -0.846 | Difficult | Requires high skills |

The results of the Rasch Partial Credit Model (PCM) analysis show variations in the levels of difficulty and ease across steps within each item. The eta parameter represents the level of difficulty of the step, while the beta parameter represents the level of ease of the step. A negative eta value and a positive beta value indicate a relatively easy step, while a positive eta value and a negative beta value indicate a difficult step (Arjana et al., 2025; Micoloi et al., 2025).

Table 3 shows strong consistency between the eta and beta parameters, which is important evidence of the psychometric quality of the instrument (Boone & Staver, 2020; Davis & Boone, 2021). The easiest steps in items P7, P8, P3, P1, P12, and P6 indicate that high school students move relatively easily to higher response categories, so these items serve as stable anchors for measuring low to mid-level critical reasoning (Prasetya & Pratama, 2023; Sujatmika et al., 2025). In contrast, the most difficult steps in items P13, P14, and P15 (especially c3) indicate that moving to the highest category requires a much higher level of critical reasoning, making these items effective discriminators (Suwita et al., 2024; Wind, 2023).

This clear variation in step difficulty and easiness supports the instrument's discriminatory power in measuring high school students' critical reasoning on the virus material (Laliyo et al., 2022; Sari & Rakhmawati, 2024). This finding aligns with previous research, which shows that Rasch instruments with varying step patterns can differentiate students' abilities hierarchically and invariantly (Fitrianto & Hidayat, 2024; Khan et al., 2023). However, this pattern also underscores the need to further evaluate difficult items to be more sensitive to aspects of reflection and argument evaluation (Kassiavera et al., 2024; Micoloi et al., 2025). Overall, the consistency between eta and beta indicates that the instrument shows good variation in the difficulty of its steps, thereby supporting its psychometric quality for measuring critical reasoning in in-depth biology learning (Arjana et al., 2025). This pattern supports Rasch's theory that instruments with clear anchors and discriminators will be more effective in supporting evidence-based learning evaluation (Boone & Staver, 2020; Wind, 2023).

Threshold analysis in the Rasch Partial Credit Model (PCM) provides important information regarding the overall item difficulty and the function of the response categories. The item location parameters range from -0.256 to 0.488, indicating a fairly good variation in difficulty between items.

**Table 4**
Location and Threshold Parameters for Each Item in the Partial Credit Rasch Model.

| Item | Location | Threshold 1 | Threshold 2 | Threshold 3 | Pola Threshold | Difficulty Level |
|---|---|---|---|---|---|---|
| P1 | -0.256 | -0.373 | -0.060 | -0.336 | Disordered | Easy |
| P2 | -0.050 | -0.705 | 0.608 | -0.052 | Disordered | Easy |
| P3 | -0.061 | -0.834 | 0.671 | -0.020 | Disordered | Easy |
| P4 | 0.239 | 0.018 | 0.568 | 0.130 | Disordered | Difficult |
| P5 | 0.008 | -0.012 | -0.332 | 0.368 | Disordered | Medium |
| P6 | -0.172 | -0.361 | -0.360 | 0.206 | Organized | Easy |
| P7 | -0.153 | -0.622 | -0.442 | 0.604 | Organized | Easy |
| P8 | -0.060 | -0.959 | 0.332 | 0.446 | Organized | Easy |
| P9 | -0.116 | -0.477 | 0.250 | -0.121 | Disordered | Easy |
| P10 | 0.334 | -0.529 | 1.057 | 0.473 | Disordered | Difficult |
| P11 | 0.189 | 0.376 | 0.076 | 0.115 | Disordered | Difficult |
| P12 | -0.241 | 0.753 | -1.365 | -0.110 | Disordered | Easy |
| P13 | 0.488 | 0.989 | -0.241 | 0.718 | Disordered | Very Difficult |
| P14 | 0.348 | 0.048 | 0.798 | 0.200 | Disordered | Difficult |
| P15 | 0.435 | 0.534 | -0.050 | 0.822 | Disordered | Difficult |

Table 4, shows that the instrument has adequate variation in item difficulty levels. Items with negative locations (P1, P6, P7, P8, P9, and P12) tend to be easier, allowing students to reach higher response categories more easily. Conversely, items with positive locations (P4, P10, P11, P13, P14, and P15) tend to be more difficult, especially item P13, which has the highest location (0.488). Despite the good variation in item locations, threshold sequence analysis revealed significant weaknesses. Only items P6, P7, and P8 showed an orderly threshold pattern (monotonically increasing). Most other items experienced disordered thresholds, meaning the threshold sequence was not sequential. This condition indicates that students tend not to distinguish the 0–3 response categories consistently, resulting in functional overlap between categories. As a result, the response categories do not fully represent a clear and gradual increase in the level of critical reasoning. The findings of disordered thresholds are important recommendations for improving the instrument in the next stage, including through improving the wording of items, providing a more explicit explanation of the meaning of each category, or considering combining response categories to make the measurement structure more stable and interpretive (Boone & Staver, 2020; Davis & Boone, 2021; Wind, 2023).

Figure 2. The results of the PCM threshold-hold analysis on the latent dimension provide a comprehensive visualization of the distribution of item difficulty and the function of the instrument's response categories. Black dots represent item locations (the overall difficulty level of the item), while white circles indicate the thresholds for moving between response categories (0 to 1, 1 to 2, and 2 to 3). The order of items from top to bottom in the plot reflects their relative difficulty. Items at the top (P1,

P12, P6, and P7) have more negative locations, so they tend to be easier. Conversely, items at the bottom (P14, P15, and especially P13) have more positive locations, which tend to make them more difficult. This pattern is highly consistent with the results of the previous descriptive statistics analysis, eta parameters, and beta (Boone & Staver, 2020; Prasetya & Pratama, 2023; Sujatmika et al., 2025). The small-person distribution around the center of the scale indicates that the majority of students have moderate critical reasoning abilities. This indicates that the overall item difficulty level is quite appropriate for the average respondent's ability, although coverage at the extremes (very low and very high ability) is still limited. (Laliyo et al., 2022; Suwita et al., 2024; Wind, 2023). However, this plot also highlights a significant structural problem with the response category function. Ideally, all three thresholds on each item should move monotonically upward from left to right as the latent dimension increases. (Boone & Staver, 2020; Khan et al., 2023; Wind, 2023). In this plot, many items, particularly P12, P13, P15, P10, and P14, exhibit overlapping or disordered threshold patterns. This indicates that students do not consistently differentiate between answer categories 0–3, leading to overlapping functions across categories. (Arjana et al., 2025; Kassiavera et al., 2024; Micoloi et al., 2025; Sari & Rakhmawati, 2024). In contrast, items such as P6, P7, and part of P8 appear more stable because their thresholds are relatively regular. Overall, the threshold plots confirm that the instrument shows good variation in item difficulty but still has weaknesses in its response-category structure. Widespread disordered thresholds are a key recommendation for future instrument revisions, such as refining item wording, more explicit category explanations, or combining response categories to improve the accuracy and stability of critical reasoning measurement (Fitrianto & Hidayat, 2024; Kwon et al., 2025; Tesio et al., 2024; Wind, 2023).

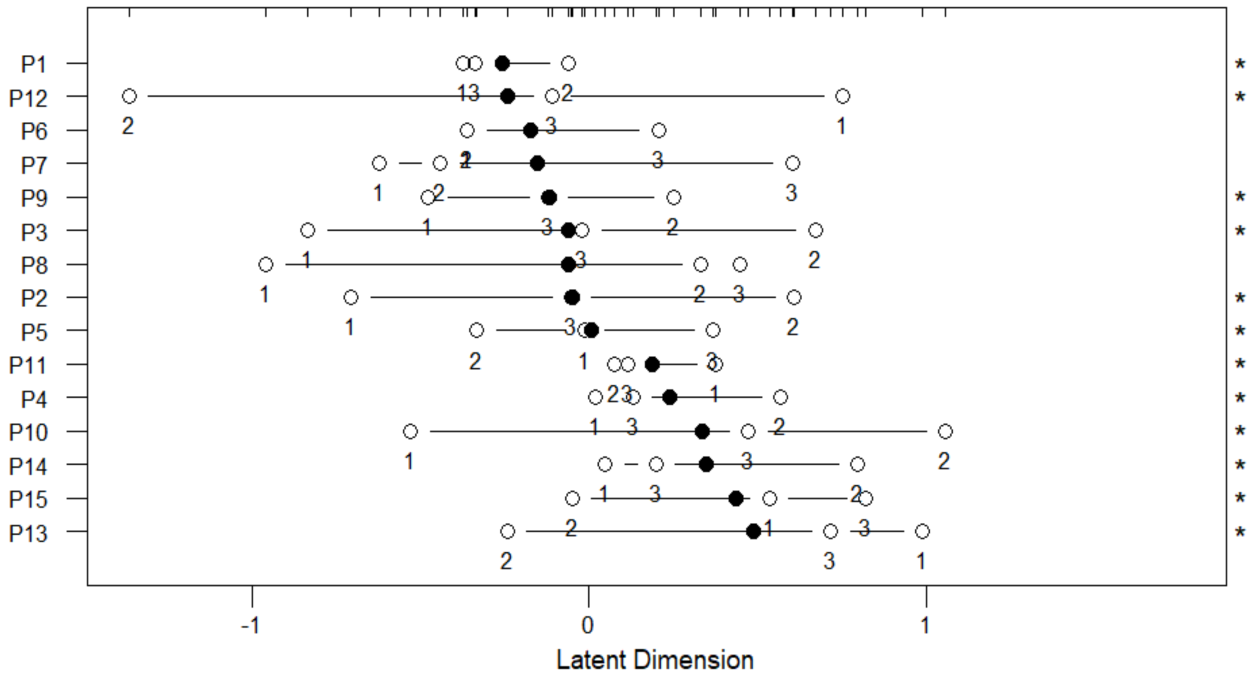


**Figure 2.** Wright Map (Person-Item Map) or Threshold Plot on Latent Dimension.

To strengthen the interpretation of the function of the response categories in the Partial Credit Model (PCM), the Item Characteristic Curve (ICC) of several items representing different patterns, namely P6, P7, P12, and P13, is displayed. The selection of these four items is based on the results of previous analyses, in which P6 and P7 tend to represent relatively easier and more stable items, while P12 and P13 represent items with a less regular threshold pattern and a more pronounced level of difficulty. Through the ICC, it can be observed how the probability of selecting each response category changes along the latent dimension, allowing assessment of whether categories 0, 1, 2, and 3 have functioned gradually and consistently.

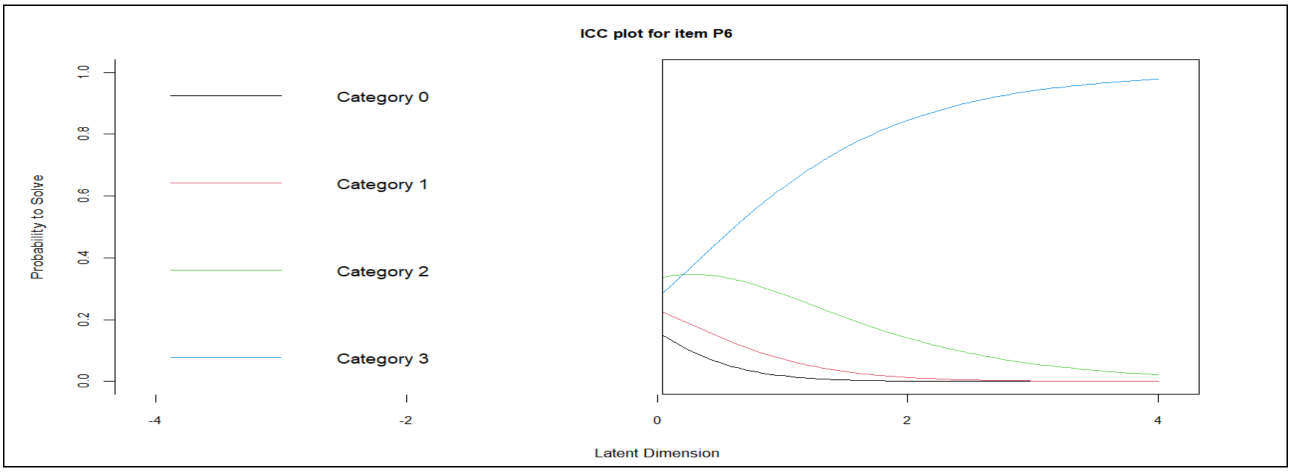


**Figure 3a.** Item Characteristic Curve (ICC) P6

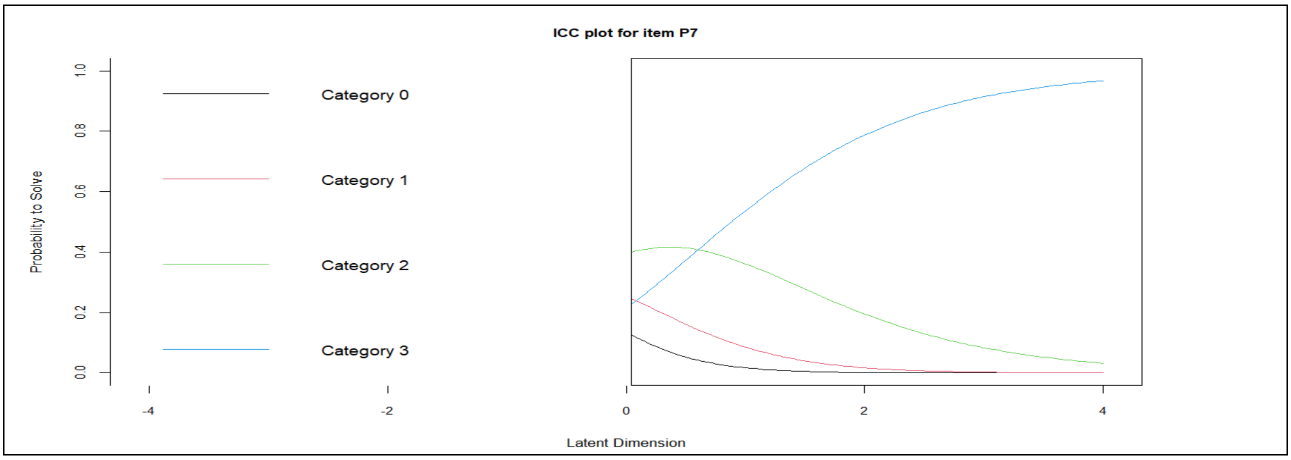


**Figure 3b.** Item Characteristic Curve (ICC) P7

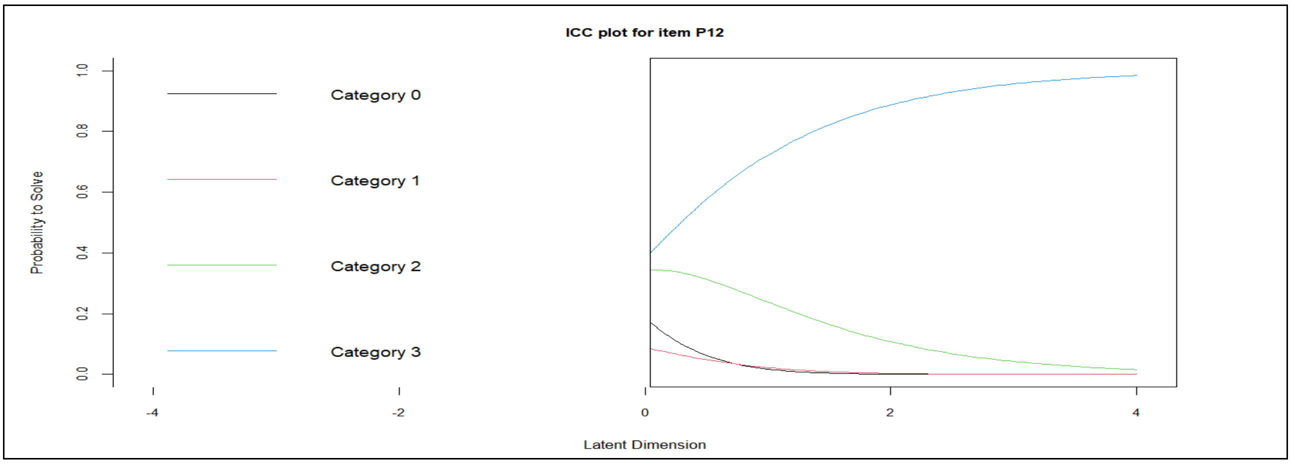


**Figure 3c.** Item Characteristic Curve (ICC) P12

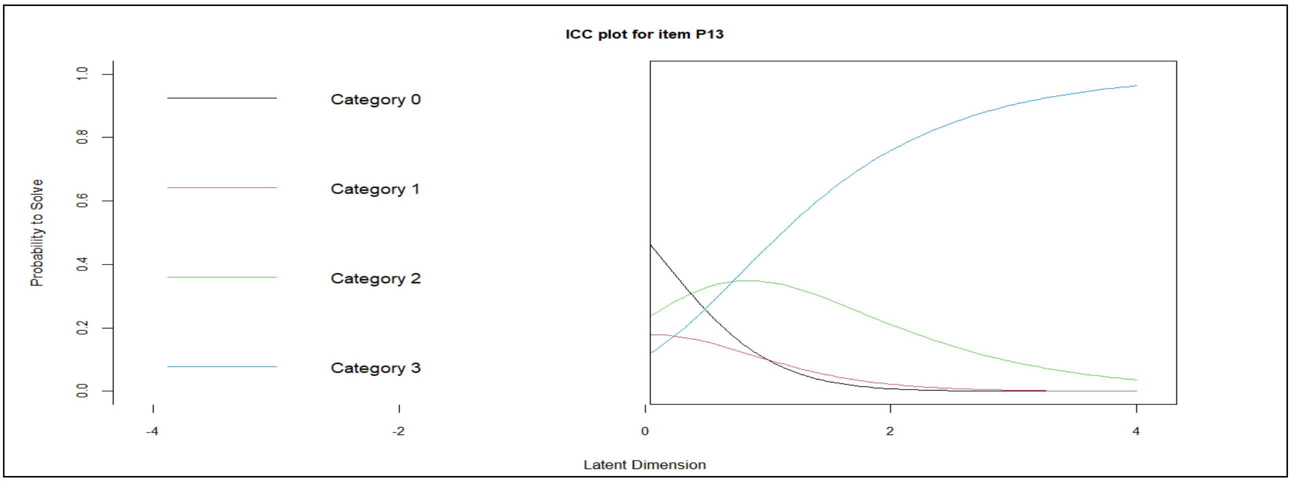


**Figure 3d.** Item Characteristic Curve (ICC) P13

Based on Figure 3a and Figure 3b. items P6 and P7 show a relatively better categorical pattern than the other items, as there is a tendency for probabilities to change between categories as the latent dimension increases. However, in these two items, the highest category (category 3) appears to begin to dominate quite quickly, leaving the middle categories, particularly category 1, functioning only within a narrow range. This indicates that although the category order in P6 and P7 is still quite legible, the discrimination between categories is not yet fully established. Figure 3c. In item P12, the curve shows that category 3 dominates across almost the entire range of the latent dimension, while category 1 has a very low presence, and category 2 appears only to a limited extent. This pattern indicates that the response categories in item P12 are not functioning optimally, as respondents tend to move directly to the highest category without making balanced use of the intermediate categories. Meanwhile, Figure 3d. item P13 exhibits a slightly more informative pattern, as category 0 remains dominant at low ability levels, category 2 peaks in the middle range, and category 3 increases at higher ability levels. Nevertheless, category 1 remains weak, so overall, the ICC results support previous findings that some response categories, particularly the middle category, still need to be evaluated to function more clearly and consistently in differentiating respondents' ability levels.

Substantively, in biology learning, viruses are characterized by highly abstract and microscopic concepts. Question structure phenomena indicate that students tend to experience a dichotomous cognitive leap. This phenomenon occurs because students either do not understand the concept at all (complete misconception, score 0) or are immediately able to construct complete reasoning (score 2 or 3) after reading the question stimulus (Romine et al., 2015). The dysfunction of the middle category proves that the assessment rubric for layered questions is less sensitive in separating students with partial understanding from those with pure misconceptions (Wetzel & Carstensen, 2014). Students experience confusion when choosing answer options because the operational boundaries between options are not clear (Kwon et al., 2025).

Overall, the PCM results indicate that the instrument exhibits a fairly good variation in step difficulty levels. Items P13, P14, and P15 tend to be the most difficult, especially in the high category, making them more selective in distinguishing respondents with higher levels of ability or latent traits. Conversely, items P7, P8, P6, P3, and P2 tend to be easier, making them more readily endorsed by most respondents. Furthermore, several items, such as P10 and especially P12, exhibit an inconsistent step pattern, necessitating further evaluation of the function of the response categories. Thus, this instrument demonstrates informative variation in difficulty, but some items still require further review in terms of their category structure. These results are further supported by the threshold analysis and ICC visualization, which indicate that although the variation in item difficulty is well-formed, the function of the response categories is not yet fully optimal. In several items, particularly P12 and P13, disordered thresholds are observed, indicating that movement between categories does not occur gradually as latent dimensionality increases. This is also reflected in the ICC curves, where certain categories, particularly the middle category, such as Category 1, do not have a clear probability peak or function only within a very narrow range. This condition indicates that respondents do not consistently distinguish among response categories, resulting in overlap between categories. Meanwhile, in relatively more stable items, such as P6 and P7, the ICC pattern shows a more systematic change in category probability as latent dimensionality increases. However, even in these items, the highest category tends to dominate more quickly, so the role of the middle category remains suboptimal. This indicates that although some items have demonstrated improved category function, the overall instrument still faces challenges in ensuring that each response category truly represents a distinct level of the construct.

The implication of these findings is the need to review the design of response categories, both in terms of the number of categories and the clarity of their descriptions. One alternative that could be considered is merging categories that are empirically too close together or simplifying the response scale to make it easier for respondents to understand. Furthermore, improving item wording could also help improve the clarity of respondents' interpretation of the categories. Thus, although this instrument has provided useful information on variations in item difficulty levels, optimizing the response categories is still needed to improve the overall quality of the measurement.

## CONCLUSION

This study successfully validated the critical reasoning assessment instrument about viruses using Winsteps software based on the Rasch Partial Credit Model, which theoretically strengthens the effectiveness of this model in measuring higher-order thinking skills in Biology learning. The analysis results show that the quality of item fit is in the adequate category because most items fit the Rasch model based on the MNSQ Infit/Outfit value, although there are irregular thresholds (eta) in most items indicating the need for simplification of response categories. The distribution of item difficulty parameters (β) shows good location variation, with items P13, P14, and P15 as the most difficult, and P6, P7, and P8 as the easiest as mapped on the Wright Map. Furthermore, the instrument separation reliability index was recorded at 0.7677, which is in the fairly good category and is able to distinguish students' abilities consistently. The practical implication of this study is the availability of potential formative diagnostic tools to support differentiated learning and deep learning according to the Curriculum. However, due to sample limitations (105 students in one district), the clinical/broad implications of the instrument require caution in generalization, so that improvements in the quality of the items and testing on larger, more diverse, and longitudinal samples are highly recommended before this instrument is implemented en masse in the evaluation of science education in Indonesia.

## ACKNOWLEDGMENT

The author would like to thank the Institute for Research and Community Service (LPPM) of Ahmad Dahlan University for funding internal research through Contract Number PD-155/SP3/LPPM-UAD/XII/2025, which enabled the successful implementation of this research.

## REFERENCES

Aggraini, M. D., Koto, I., & Gunawan, B. (2024). The Effect of Problem-Based Learning Integrated Numbered Head Together on the Critical Thinking Ability of Grade XI Students. *Kasuari: Physics Education ….* https://journalfkipunipa.org/index.php/kpej/article/view/780

Akker, J. van den, Bannan, B., Kelly, A. E., Nieveen, N., & Plomp, T. (2013). Educational Design Research Part A: An introduction. In T. Plomp & N. Nieveen (Eds.), *Educational design research. Part A : An introduction* (1st ed.). Netherlands Institute for Curriculum Development (SLO), Enschede, the Netherlands.

Alnahdi, A. H., Alsubiheen, A. M., & Aldaihan, M. M. (2025). Rasch Measurement Model Supports the Unidimensionality and Internal Structure of the Arabic Oswestry Disability Index. *Journal of Clinical Medicine*, *14*(4), 1259. https://doi.org/10.3390/jcm14041259

Aristizábal Gómez, Y. M., Jiménez Sierra, Á. A., & Ortega Iglesias, J. M. (2025). Students' Emotions Toward Assessments: A Systematic Review. *Social Sciences*, *14*(11), 652. https://doi.org/10.3390/socsci14110652

Arjana, I. M., Parmiti, D. P., Candiasa, I. M., & Widiartini, N. K. (2025). Unlocking the predictive power of the Rasch model: A systematic literature review on educational instrument calibration and assessment accuracy. *Edelweiss Applied Science and Technology*, *9*(6), 2003–2022. https://doi.org/10.55214/25768484.v9i6.8310

Boone, W. J., & Staver, J. R. (2020). Advances in Rasch Analyses in the Human Sciences. In *Advances in Rasch Analyses in the Human Sciences*. Springer International Publishing. https://doi.org/10.1007/978-3-030-43420-5
BSKP. (2022). Panduan Pembelajaran dan Asesmen Pendidikan Anak Usia Dini pada Kurikulum Merdeka. In *Kemendikbudristek* (p. 123). Kepala Badan Standar, Kurikulum, dan Asesmen Pendidikan Kementerian Pendidikan Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi. https://repositori.kemendikdasmen.go.id/33290/
Calma, A., & Davies, M. (2025). Assessing students' critical thinking abilities via a systematic evaluation of essays. *Studies in Higher Education*, *51*(2), 422–437. https://doi.org/10.1080/03075079.2025.2470969
Davis, D. R., & Boone, W. (2021). Using Rasch analysis to evaluate the psychometric functioning of the other-directed, lighthearted, intellectual, and whimsical (OLIW) adult playfulness scale. *International Journal of Educational Research Open*, *2*, 100054. https://doi.org/10.1016/j.ijedro.2021.100054
Eneng Sururiyatul Mu'aziyah, S., Edy Nurtamam, M., & Nurmegawati, L. (2025). Mapping secondary school students' botanical literacy using rasch analysis to support deep learning in science. *EPJ Web of Conferences*, *344*, 01038. https://doi.org/10.1051/epjconf/202534401038
Fitrah, M., Sofroniou, A., Ofianto, Judijanto, L., & Widihastuti. (2024). Reliability and separation index analysis of mathematics questions integrated with the cultural architecture framework using the Rasch model. *Journal of Education and E-Learning Research*, *11*(3), 499–509. https://doi.org/10.20448/jeelr.v11i3.5861
Fu, R., & Koh, K. (2026). Developing and validating teacher formative assessment literacy questionnaire in the high-stakes examination culture: a case of China. *Frontiers in Psychology*, *17*. https://doi.org/10.3389/fpsyg.2026.1771941
Golden, B. (2023). Enabling critical thinking development in higher education through the use of a structured planning tool. *Irish Educational Studies*, *42*(4), 949–969. https://doi.org/10.1080/03323315.2023.2258497
Heim, A. B., Esparza, D., Holmes, N. G., & Smith, M. K. (2023). Comparing study features is easy but identifying next steps is hard: Evaluating critical thinking through the Biology Lab Inventory of Critical Thinking in Ecology. *Ecology and Evolution*, *13*(5). https://doi.org/10.1002/ece3.10071
Hidayati, N., Suryanti, S., & Cahaya, N. (2023). The Potential for the Development of Critical Thinking Tests : An Overview of Educators' Perceptions. *Jurnal Kependidikan: Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan, Pengajaran Dan Pembelajaran*, *9*(1), 56. https://doi.org/10.33394/jk.v9i1.6794
Hu, X., & Bi, H. (2025). Exploring and validating the componential model of students' scientific critical thinking in science education. *Thinking Skills and Creativity*, *55*, 101695. https://doi.org/10.1016/j.tsc.2024.101695
Ibnu Fitrianto, & Al Mukaromah Hidayat. (2024). Critical Reasoning Skills: Designing an Education Curriculum Relevant to Social and Economic Needs. *International Journal of Post Axial: Futuristic Teaching and Learning*, 245–258. https://doi.org/10.59944/postaxial.v2i4.393
Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia*, *12*(1), 40. https://doi.org/10.1186/s40468-022-00191-4
Kassiavera, S., Suparmi, A., Cari, C., & Sukarmin, S. (2024). Application of Rasch Model in Two-Tier Test for Assessing Critical Thinking in Physics Education. *Journal of Baltic Science Education*, *23*(6), 1227–1242. https://doi.org/10.33225/jbse/24.23.1227
Khan, M. S. H., Salele, N., Hasan, M., & Abdou, B. O. (2023). Factors affecting student readiness towards OBE implementation in engineering education: Evidence from a developing country. *Heliyon*, *9*(10), e20905. https://doi.org/10.1016/j.heliyon.2023.e20905
Kusurkar, R. A., Orsini, C., Somra, S., Artino, A. R., Daelmans, H. E. M., Schoonmade, L. J., & Van Der Vleuten, C. (2023). The Effect of Assessments on Student Motivation for Learning and Its Outcomes in Health Professions Education: A Review and Realist Synthesis. *Academic Medicine*, *98*(9), 1083–1092. https://doi.org/10.1097/ACM.0000000000005263
Kwon, J., Raghunandan, R., Nghiem, S. H., Howard, K., Lancsar, E., Huynh, E., Howell, M., Petrou, S., & Smith, S. (2025). *Rasch analysis of the self-reported PedsQL ™ 4 . 0 Generic Core Scales by Australian*

*children*. *4*.
Laliyo, L. A. R., Sumintono, B., & Panigoro, C. (2022). Measuring changes in hydrolysis concept of students taught by inquiry model: stacking and racking analysis techniques in Rasch model. *Heliyon*, *8*(3), e09126. https://doi.org/10.1016/j.heliyon.2022.e09126
Mallinson, T., Kozlowski, A. J., Johnston, M. V., Weaver, J., Terhorst, L., Grampurohit, N., Juengst, S., Ehrlich-Jones, L., Heinemann, A. W., Melvin, J., Sood, P., & Van de Winckel, A. (2022). Rasch Reporting Guideline for Rehabilitation Research (RULER): the RULER Statement. *Archives of Physical Medicine and Rehabilitation*, *103*(7), 1477–1486. https://doi.org/10.1016/j.apmr.2022.03.013
Micoloi, M., Ivanjek, L., De Cock, M., Schubatzky, T., Wildbichler, S., Wackermann, R., & Pospiech, G. (2025). Development of a multiple-choice test to assess critical thinking in climate change. *Physical Review Physics Education Research*, *21*(2), 020154. https://doi.org/10.1103/jnbq-1qyk
Pat-El, R. J., de Hoog, N., Segers, M., & Vedder, P. (2024). Exploring the impact of student perceptions of Assessment for Learning on intrinsic motivation. *Studies in Educational Evaluation*, *83*. https://doi.org/10.1016/j.stueduc.2024.101420
Prasetya, W. A., & Pratama, A. T. (2023). Item quality analysis using the Rasch model to measure critical thinking ability in the material of the human digestive system of Biology subject in high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *27*(1), 76–91. https://doi.org/10.21831/pep.v27i1.58873
Retnawati, H., Hadi, S., & Nugraha, A. C. (2016). Vocational high school teachers' difficulties in implementing the assessment in curriculum 2013 in Yogyakarta Province of Indonesia. *International Journal of Instruction*, *9*(1), 33–48. https://doi.org/10.12973/iji.2016.914a
Rodríguez Rincón, Y., Munárriz, A., & Magreñán Ruiz, A. (2024). A new approach to continuous assessment: Moving from a stressful sum of grades to meaningful learning through self-reflection. *International Journal of Management Education*, *22*(3), 101072. https://doi.org/10.1016/j.ijme.2024.101072
Romine, W. L., Schaffer, D. L., & Barrow, L. (2015). Development and Application of a Novel Rasch-based Methodology for Evaluating Multi-Tiered Assessment Instruments: Validation and utilization of an undergraduate diagnostic test of the water cycle. *International Journal of Science Education*, *37*(16), 2740–2768. https://doi.org/10.1080/09500693.2015.1105398
Sari, T. N. I., & Rakhmawati, A. (2024). Analysis the quality of critical thinking and creativity questions in high school biology subjects with the Rasch model. *Research and Development in Education (RaDEn)*, *4*(1), 602–616. https://doi.org/10.22219/raden.v4i1.32758
Sartono, B., Sunarno, W., Prayitno, B. A., & Indriyanti, N. Y. (2025). Development and Validation of A Critical Thinking Assessment on Temperature and Heat for Secondary Physics Education. *Educational Process: International Journal*, *16*(1). https://doi.org/10.22521/edupij.2025.16.264
Setiowati, W., Hariyadi, S., & Prihandono, T. (2025). Enhancing student piety and critical reasoning through steamrel-PBL with authentic e-assessment in ecology learning. *Biosfer*, *18*(2), 1–13. https://doi.org/10.21009/biosferjpb.50319
Siswanto, D. H., & Peni, N. R. N. (2023). Publication Trend on the Plomp Development Model in Mathematics Education. *Asian Pendidikan*, *3*, 71–80. https://doi.org/10.53797/aspen.v3i2.9.2023
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
Sutiarso, S., Rosidin, U., & Sulistiawan, A. (2022). Developing Assessment Instrument Using Polytomous Response in Mathematics. *European Journal of Educational Research*, *11*(3), 1441–1462. https://doi.org/10.12973/eu-jer.11.3.1441
Suwita, S., Saputro, S., Sajidan, S., & Sutarno, S. (2024). Assessing Lower-Secondary School Students' Critical Thinking Skills in Photosynthesis: a Rasch Model Approach. *Journal of Baltic Science Education*, *23*(6), 1278–1290. https://doi.org/10.33225/jbse/24.23.1278
Tavares, R., Vieira, R. M., & Pedro, L. (2021). Mobile app for science education: Designing the learning approach. *Education Sciences*, *11*(2), 1–23. https://doi.org/10.3390/educsci11020079
Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 1. The "most likely" measures coming from the model. *Disability and Rehabilitation*, *46*(3), 591–603.

https://doi.org/10.1080/09638288.2023.2169771
Thalhah, S. Z., Tahrim, T., & Sari, R. E. (2022). Development of Lecturer Performance Assessment Instruments through the Plomp Model in Higher Education. *AL-ISHLAH: Jurnal Pendidikan*, *14*(2), 1527–1540. https://doi.org/10.35445/alishlah.v14i2.1350
Usman, U., Repelita, R., Setiawan, R., & Hidayat, S. (2024). Differentiated Learning and Critical Thinking Skills Development in Biology: A Systematic Literature Review. *Indonesian Journal of Multidiciplinary Research*, *4*(2), 447–454. https://doi.org/10.17509/ijomr.v4i2.78729
Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation*, *83*. https://doi.org/10.1016/j.stueduc.2024.101425
Wahyuni, V., Ratnawulan, & Mufit, F. (2025). HOTS Assessment Development to Assess High School Physics Problem-Solving Skills. *Jurnal Penelitian Pendidikan IPA*, *11*(10), 536–546. https://doi.org/10.29303/jppipa.v11i10.12880
Wetzel, E., & Carstensen, C. H. (2014). Reversed Thresholds in Partial Credit Models: A Reason for Collapsing Categories? *Assessment*, *21*(6), 765–774. https://doi.org/10.1177/1073191114530775
Wind, S. A. (2023). Detecting Rating Scale Malfunctioning With the Partial Credit Model and Generalized Partial Credit Model. *Educational and Psychological Measurement*, *83*(5), 953–983. https://doi.org/10.1177/00131644221116292
Wuisan, A. C., Simamora, M., & Siboro, M. (2024). Efektifitas Problem Based Learning Terhadap Kemampuan Berpikir Kritis Siswa [The Effectiveness of Problem Based Learning on Students' Critical Thinking Skills]. *Jurnal Teropong Pendidikan*. http://ojs.uph.edu/index.php/JTP/article/view/8638
Yang, J., Reheman, Z., Liu, Y., & Zhao, S. (2023). Applying Rasch analysis in refinement and validation of interpersonal skills measure for gifted children. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1236640
Yokhebed, Karmadi, R. M. D., Silaen, J., Meidina, S., & Nikmah, N. A. (2025). Analysis of the Quality of Sudents' Critical Reasoning Instruments Using Rasch Model. *Jurnal Penelitian Pendidikan IPA*, *11*(3), 54–60. https://doi.org/10.29303/jppipa.v11i3.9584
Yokhebed, Y., Karmadi, R. M. D., & Nastiti, L. R. (2025). Validity and reliability analysis of a socioscientific issues-based critical thinking self-assessment instrument using the Rasch model. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *11*(1), 73–82. https://doi.org/10.22219/jpbi.v11i1.38902
Zhao, H., Yu, J., & Li, G. (2026). Developing a process-oriented classroom observation protocol for assessing high school students' computational thinking in science classrooms: a Rasch-based proficiency framework. *Frontiers in Psychology*, *17*. https://doi.org/10.3389/fpsyg.2026.1722422