

Psychometric Evaluation of a Critical Thinking Skills Instrument on Renewable Energy Using the Rasch Model

Rapi Alwi Azhar*, Parlindungan Sinaga, Agus Danawan, Mohd Khairul Nizam Syah, Lina Aviyanti

Physics Education, Faculty of Mathematics and Natural Science Education, Universitas Pendidikan Indonesia, Indonesia

*Correspondence to: rapiawiazhar@upi.edu

Abstract: Critical thinking is an essential twenty-first-century skill that enables students to analyze information, evaluate evidence, and make reasoned decisions. Therefore, a high-quality instrument is needed to accurately measure students' critical thinking skills. This study aimed to analyze the quality of a critical thinking skills instrument on renewable energy using the Rasch model. The study employed a descriptive quantitative design involving 75 tenth-grade students. The instrument consisted of 15 multiple-choice items developed based on Ennis's critical thinking framework, including basic clarification, bases for a decision, inference, advanced clarification, and supposition and integration. Data were analyzed using the Rasch model through construct validity, reliability, item difficulty, Differential Item Functioning (DIF), and Wright Map analyses. The results showed that the instrument satisfied the unidimensionality requirement, with a raw variance explained by measures of 28.8%. Most items demonstrated acceptable fit to the Rasch model, although one item showed misfit and another required revision. The instrument achieved fair person reliability (0.67), fair item reliability (0.78), and very good internal consistency (Cronbach's $\alpha = 0.79$). The DIF analysis indicated that all items were free from significant gender bias. Furthermore, the Wright Map revealed a mistargeting condition in which the average student ability exceeded the average item difficulty. Overall, the instrument demonstrated acceptable measurement quality and can be used to assess students' critical thinking skills on renewable energy topics, although several items require further refinement.

Keywords: Critical thinking skills; Renewable energy; Rasch model; Instrument quality; Physics education

Introduction

Critical thinking is one of the essential twenty-first-century competencies that students need to address increasingly complex global challenges (Facione, 2020; OECD, 2019). In Indonesia, the importance of developing critical thinking skills is reflected in the Merdeka Curriculum, which explicitly promotes Higher Order Thinking Skills (HOTS)-based learning across all subjects, including physics (Kemendikbudristek, 2022). This emphasis is consistent with the national educational goals of fostering individuals who are knowledgeable, competent, creative, and independent. Within physics education, renewable energy serves as a particularly relevant context for developing critical thinking because it involves multidimensional issues such as technological efficiency, environmental impacts, and energy policies that require careful analysis and evaluation (Bao & Koenig, 2019; Care et al., 2018). The contextual and decision-oriented nature of renewable energy topics makes them suitable for assessing students' critical thinking skills, as demonstrated in prior studies measuring students' critical thinking on renewable energy and related physics topics (Febriani et al., 2023; Yuliarti et al., 2023; Mahir & Martawijaya, 2025). Therefore, instruments are needed to measure students' abilities to analyze information, evaluate alternative solutions, and draw evidence-based conclusions.

According to (Ennis, 2011), critical thinking consists of five major dimensions: basic clarification, bases for a decision, inference, advanced clarification, and supposition and integration. Compared to other well-established frameworks such as Facione (1990) APA Delphi model, which emphasizes

dispositions alongside cognitive skills, or the Paul & Elder (2008) model, which offers discipline-neutral standards of reasoning, Ennis's framework was selected because its dimensions are more readily operationalized into discrete, measurable cognitive indicators and have already been translated into validated instruments (e.g., the Cornell Critical Thinking Test) widely applied in science and physics education research. These dimensions cannot be adequately measured through simple recall-based assessments; instead, they require carefully designed instruments that elicit higher-order cognitive processes. Consequently, the development of valid and reliable critical thinking assessment instruments is a crucial prerequisite before such instruments can be used in experimental studies or educational evaluations (Aiken, 1985; Retnawati, 2016).

Despite their importance, the quality of existing critical thinking assessment instruments remains a significant concern. Several studies have reported that many instruments lack sufficient empirical testing of essential psychometric properties including content and construct validity, reliability, and item-level indices such as difficulty and discrimination making their overall measurement quality questionable (Ardiansyah et al., 2022; Maulana et al., 2020; Sujatmika et al., 2025). Furthermore, some instruments fail to identify item bias, specifically differential item functioning (DIF) across different groups, or exhibit a mismatch between item difficulty and students' ability distributions. These issues may lead to inaccurate measurement results and misleading interpretations of learning outcomes.

Empirical evaluation of instrument quality is commonly conducted using either Classical Test Theory (CTT) or Item Response Theory (IRT), one of which is the Rasch model. A major limitation of CTT is that item parameters, such as difficulty and discrimination indices, are sample-dependent, meaning that their values may vary according to the characteristics of the sample being analyzed (Bond & Fox, 2007; Bond et al., 2020). In contrast, the Rasch model estimates student ability (person measure) and item difficulty (item measure) on the same logit scale and provides sample-independent measurement, allowing for more objective and stable comparisons across groups and contexts. This property makes the Rasch model particularly appropriate for the present study: by placing item difficulty and student ability on a common interval scale, it directly addresses the two psychometric problems identified earlier namely, the difficulty in detecting differential item functioning (DIF) and the mismatch between item difficulty and students' ability distributions thereby supporting the development of a critical thinking instrument that is both valid and invariant across student groups.

The advantages of the Rasch model extend beyond sample independence. Rasch analysis provides extensive diagnostic information that enables researchers to evaluate instrument unidimensionality as evidence of construct validity, examine item fit using outfit mean square (MNSQ) and standardized fit statistics (ZSTD), identify potential item bias through Differential Item Functioning (DIF), and visualize the relationship between student ability and item difficulty using a Wright Map (Sumintono & Widhiarso, 2015). These advantages have made the Rasch model one of the most widely recommended approaches for developing high-quality educational assessment instruments (Alias & Hafiz, 2021; Aviyanti et al., 2023).

Although the Rasch model has been extensively employed to evaluate educational instruments measuring concept understanding, scientific literacy, and scientific attitudes (Boone et al., 2014), its application to critical thinking assessment in physics education has only recently begun to receive attention (Bhakti et al., 2023; Kassiavera et al., 2024; Halim et al., 2026). Existing studies have demonstrated the usefulness of Rasch analysis for establishing instrument validity and reliability; however, most have emphasized selected psychometric properties rather than providing an integrated evaluation encompassing unidimensionality, item fit, reliability, Differential Item Functioning (DIF), item targeting, and Wright Map interpretation. Furthermore, studies applying such comprehensive Rasch analyses to critical thinking instruments in the specific context of renewable energy remain scarce. Therefore, a more comprehensive psychometric evaluation is needed to strengthen the evidence supporting the quality of critical thinking assessment instruments in this domain.

Therefore, this study aims to conduct a comprehensive quality analysis of a critical thinking skills

instrument on renewable energy using the Rasch model. Unlike previous studies that primarily reported internal reliability alone, this study integrates analyses of unidimensionality, item fit, reliability, item difficulty, Differential Item Functioning (DIF), and Wright Map interpretation to provide a more comprehensive evaluation of the instrument's quality.

Methods

Research Design

This study employed a descriptive quantitative design focusing on the evaluation of instrument quality using the Rasch model. The instrument analyzed was a multiple-choice critical thinking skills test consisting of 15 items on renewable energy. The indicators of critical thinking skills were developed based on Ennis (2011) framework, which includes basic clarification, bases for a decision, inference, advanced clarification, and supposition and integration. The five indicators were distributed across the 15 items, with each indicator corresponding to a set of item numbers. The distribution of indicators and item numbers is presented in Table 1.

Table 1. Distribution of Critical Thinking Indicators, Test Items, and Their Characteristics

Indicator	Item number	Item Characteristics
Basic Clarification	S1, S2, S3	Identify relevant information, interpret contextual problems, and analyze arguments related to renewable energy.
Base for a Decision	S4, S5, S6	Evaluate the credibility of information sources and observation reports to support evidence-based decisions.
Inference	S7, S8, S9	Draw logical conclusions from statements, graphs, and data using deductive or inductive reasoning.
Advance Clarification	S10, S11	Evaluate scientific definitions and distinguish accurate concepts from misconceptions.
Supposition and Integration	S12, S13, S14, S15	Predict logical consequences and justify decisions by integrating scientific evidence and sustainability considerations.

Participants

The study involved 75 tenth-grade students who had completed learning activities related to renewable energy. The participants consisted of 26 male students and 49 female students selected using purposive sampling. This sample size is considered adequate for Rasch model analysis, as Peabody et al. (2020) recommends a minimum of 30 participants to obtain stable item calibration, while Bond & Fox (2007) affirm that samples between 50 and 100 are sufficient for reliable parameter estimation in dichotomous Rasch models, particularly when the number of items is moderate.

Data Collection

Student responses to each item were scored dichotomously, with a score of 1 assigned for a correct answer and 0 for an incorrect answer. This scoring approach is consistent with the requirements of the dichotomous Rasch model, which assumes that each item has only two possible response categories (Bond & Fox 2007; Sumintono & Widhiarso, 2015). The use of dichotomous scoring ensures

that the data structure aligns with the model's assumptions, thereby supporting the validity of the subsequent Rasch-based analyses.

Data Analysis

The quality of the instrument was evaluated using the Rasch model through MINISTEP version 5.10.3.0. The analysis included construct validity, reliability, item difficulty, Differential Item Functioning (DIF), and Wright Map analysis.

Construct validity was examined through unidimensionality and item fit statistics. Unidimensionality was evaluated using the raw variance explained by measures, which indicates the proportion of variance accounted for by the measured construct. Item fit was assessed using three indicators: outfit mean square (MNSQ), outfit standardized fit statistic (ZSTD), and point-measure correlation (Pt. Measure Corr). An item is considered to fit the Rasch model when MNSQ falls between 0.5 and 1.5, ZSTD between -2.0 and +2.0, and Point Measure Correlation between 0.4 and 0.85. The criteria used for evaluating construct validity are presented in Table 2.

Table 2. Criteria for Unidimensionality and Item Fit

Analysis Aspect	Indicator	Criterion
Unidimensionality	Raw variance explained by measures	>20% = acceptable
		>40% = good
		>60% = excellent
Item Fit	Outfit MNSQ	0,5 < MNSQ < 1,5
	Outfit ZSTD	-2,0 < ZSTD < +2,0
	Point Measure Corr	0,4 < Point Measure Corr < 0,85

(Bond & Fox, 2007; Sumintono & Widhiarso, 2015)

Reliability was evaluated using person reliability, item reliability, and Cronbach's alpha (KR-20). Person reliability indicates the consistency of students' responses, whereas item reliability reflects the stability of item difficulty estimates. Cronbach's alpha was used to assess the internal consistency of the instrument. Both person and item reliability were classified into five categories ranging from poor (<0.67) to excellent (>0.94), while Cronbach's alpha ranged from poor (<0.50) to excellent (>0.80). The interpretation criteria are presented in Table 3.

Table 3. Reliability Interpretation Criteria

Reliability Index	Value Range	Category
Pearson & Item reliability	< 0,67	Poor
	0,67 – 0,80	Fair
	0,81 – 0,90	Good
	0,91 – 0,94	Very Good
	> 0,94	Excellent
Cronbach Alpha (KR – 20)	< 0,50	Poor
	0,50 – 0,60	Fair
	0,61 – 0,70	Good
	0,71 – 0,80	Very Good
	> 0,80	Excellent

(Bond & Fox, 2007; Sumintono & Widhiarso, 2015)

Item difficulty was estimated using the Joint Maximum Likelihood Estimation (JMLE) measure expressed in logits. Item difficulty categories were determined relative to the mean item measure and

standard deviation, classifying items into four levels: very easy, easy, difficult, and very difficult. The item difficulty categories are presented in Table 4.

Table 4. Item Difficulty Categories

Logit Range	Category
$\text{Logit} < \text{Mean} - \text{SD}$	Very easy
$\text{Mean} - \text{SD} \leq \text{Logit} < \text{Mean}$	Easy
$\text{Mean} \leq \text{Logit} < \text{Mean} + \text{SD}$	Difficult
$\text{Logit} \geq \text{Mean} + \text{SD}$	Very Difficult

(Boone et al., 2014)

Differential Item Functioning (DIF) analysis was conducted to identify potential item bias based on gender. Items were considered to exhibit DIF when the Rasch-Welch probability value was less than 0.05.

In addition, a Wright Map was used to visualize the distribution of student ability and item difficulty on the same logit scale. The Wright Map enabled the examination of targeting between student ability and item difficulty, the identification of measurement gaps across different ability levels, and informed decisions regarding the revision or development of additional items to improve measurement coverage. Furthermore, the results of each Rasch analysis informed item retention decisions. Items meeting the established fit criteria were retained, whereas misfitting items were reviewed based on both statistical evidence and content considerations to determine whether revision or deletion was warranted. DIF analysis was used to identify potentially biased items requiring further evaluation, while Wright Map analysis informed the need for additional items to better represent the full range of student abilities (Bond et al., 2020; Boone et al., 2014).

Results and Discussion

Construct Validity

Unidimensionality

Construct validity was first examined through the unidimensionality assumption. The analysis revealed that the instrument met the minimum unidimensionality criterion, with a raw variance explained by measures of 28.8%, which exceeds the 20% threshold required for acceptable construct validity. The detailed results are presented in Table 5.

Table 5. Unidimensionality Analysis Results

Unidimensionality Indicator	Value	Interpretation
Raw variance explained by measures	28,8%	Acceptable

The raw variance explained by measures was 28.8%, exceeding the minimum criterion of 20% established by Sumintono & Widhiarso (2015). This result indicates that the instrument satisfies the unidimensionality assumption and that a dominant construct, namely critical thinking skills, underlies student responses across items.

However, it is important to acknowledge that 28.8% represents a relatively modest level of explained variance. In comparison, studies reporting Rasch-based instruments with stronger unidimensionality typically demonstrate values between 40% and 60% (Boone et al., 2014; Linacre, 2016). The relatively lower value obtained in this study may reflect the multidimensional nature of critical thinking itself, which encompasses distinct cognitive processes such as inference, evaluation, and advanced clarification (Ennis, 2011). While these sub-dimensions are theoretically interrelated, they may introduce moderate construct heterogeneity, thereby reducing the proportion of variance explained by a single dominant factor. Nevertheless, as the value exceeds the minimum threshold, the instrument retains acceptable construct validity and can be considered sufficiently unidimensional for the purposes of Rasch analysis. Future instrument development should consider refining item alignment across

critical thinking dimensions to enhance construct coherence and increase explained variance.

Item Fit Analysis

The item fit analysis showed that 13 of the 15 items demonstrated acceptable fit to the Rasch model. Item S7 was identified as misfit, while item S2 showed weak discrimination and requires revision. The complete item fit statistics are presented in Table 6.

Table 6. Item Fit Statistics

Butir Soal	Outfit		Pt. Measure Corr	Interpretation
	Outfit MNSQ	Outfit ZSTD		
S1	0,88	-0,50	0,54	Fit
S2	1,49	2,42	0,25	Needs revision
S3	0,78	-0,85	0,59	Fit
S4	0,78	-1,17	0,61	Fit
S5	1,00	0,09	0,52	Fit
S6	0,79	-1,07	0,60	Fit
S7	1,86	3,72	0,15	Misfit
S8	0,57	-1,03	0,58	Fit
S9	0,80	-1,02	0,59	Fit
S10	1,12	0,50	0,41	Fit
S11	1,19	0,81	0,45	Fit
S12	0,42	-1,59	0,67	Acceptable
S13	1,10	0,53	0,46	Fit
S14	0,70	-1,11	0,59	Fit
S15	0,53	-1,29	0,61	Fit

Thirteen of the fifteen items met at least two of the three fit criteria, indicating that the majority of items functioned consistently with the Rasch model and contributed meaningfully to the measurement of critical thinking skills. This proportion of fitting items is comparable to findings from similar instrument development studies in physics education, where item fit rates of 80–90% are commonly reported as acceptable (Ardiansyah et al., 2022; Aviyaniti et al., 2023).

Item S7 demonstrated substantial misfit, with an Outfit MNSQ of 1.86 and an Outfit ZSTD of 3.72, both exceeding the recommended thresholds of 0.5–1.5 and ± 2.0 , respectively. The Pt Measure Corr value of 0.15 further indicates that this item had limited ability to discriminate between students of different ability levels. According to Rasch measurement literature, item misfit may arise from ambiguous wording, ineffective distractors, or a mismatch between an item's cognitive demands and the intended construct (Bond & Fox, 2007; Halim et al., 2026). Content analysis of Item S7 supports the latter interpretation. The item requires students to interpret a graph showing renewable energy production trends in Indonesia from 2019 to 2022 and identify the conclusion best supported by the presented data. However, one response option asks students to infer future developments beyond the evidence presented in the graph, requiring predictive reasoning rather than inference based solely on the available data. This additional reasoning demand may have introduced construct-irrelevant variance, resulting in inconsistent response patterns among students with similar levels of critical thinking ability. Therefore, the item should be revised by ensuring that all response options require conclusions derived exclusively from the evidence presented in the graph, thereby aligning more closely with the intended inference construct.

Item S2 demonstrated a point-measure correlation of 0.25, which falls below the recommended minimum of 0.30 (Bond & Fox, 2007), indicating weak discrimination. Although its Outfit MNSQ of 1.49 remained within the marginal acceptable range, the low correlation suggests that the item did not function as effectively as the other items in differentiating students. This may reflect overly complex or ambiguous stem construction that confused even high-ability students. The item warrants targeted revision, particularly in clarifying its cognitive demand and ensuring alignment with the inference or evaluation dimensions of critical thinking. Item S12, while showing an Outfit MNSQ slightly outside the recommended lower bound, was retained as acceptable given that its ZSTD and point-measure correlation satisfied the established criteria.

Reliability

The reliability analysis yielded a person reliability of 0.67 (fair), item reliability of 0.78 (fair), and a Cronbach's alpha of 0.79 (very good), indicating that while internal consistency was satisfactory, measurement precision at the person level requires improvement. The detailed reliability indices are presented in Table 7.

Table 7. Reliability Analysis Results

Index	Value	Category
Person Reliability	0,67	Fair
Item Reliability	0,78	Fair
Cronbach Alpha KR-20	0,79	Very Good

The person reliability value of 0.67 indicates a fair level of consistency in student response patterns. This value suggests that the instrument was moderately effective in separating students into distinct ability strata, but lacked the precision needed to finely differentiate individuals across the full ability spectrum. Compared to high-quality critical thinking instruments that typically achieve person reliability values above 0.80 (Boone et al., 2014; Wright & Masters, 1982), the present value reflects a limitation that is likely attributable to the relatively small number of items ($n = 15$). As Linacre (1994) notes, person reliability is sensitive to test length, and instruments with fewer than 20 items often yield moderate person reliability even when items themselves are well-calibrated. Future revisions that expand the item pool, particularly at higher difficulty levels, are expected to improve this index.

The item reliability value of 0.78 indicates that the hierarchy of item difficulty was sufficiently stable and replicable across similar student populations. This finding implies that the relative ordering of item difficulty would likely be reproduced if the instrument were administered to a comparable sample, an important property for instruments intended for broader use in physics education assessment. The Cronbach's alpha coefficient of 0.79 further demonstrates good internal consistency, confirming that the items collectively measured a coherent underlying construct. Taken together, the reliability indices suggest that while the instrument demonstrates acceptable psychometric properties at the item level, further development is needed to strengthen measurement precision at the person level.

Item Difficulty

The item difficulty analysis revealed a moderate spread ranging from -1.11 to 1.09 logits, with the majority of items (eight out of fifteen) categorized as difficult or very difficult, reflecting the cognitively demanding nature of critical thinking tasks on renewable energy topics. The complete item difficulty estimates based on JMLE measures are presented in Table 8.

Table 8. Item Difficulty Analysis

Item	JMLE Measure	SD	Category
S1	0,19	0,66	Difficult
S2	0,88	0,66	Very Difficult

Item	JMLE Measure	SD	Category
S3	-0,15	0,66	Easy
S4	0,43	0,66	Difficult
S5	1,09	0,66	Very Difficult
S6	0,43	0,66	Difficult
S7	0,58	0,66	Difficult
S8	-1,11	0,66	Very Easy
S9	0,35	0,66	Difficult
S10	-0,34	0,66	Easy
S11	-0,06	0,66	Easy
S12	-1,11	0,66	Very Easy
S13	0,11	0,66	Difficult
S14	-0,34	0,66	Easy
S15	-0,98	0,66	Very Easy

The distribution of item difficulty ranged from -1.11 to 1.09 logits, spanning a moderate but somewhat narrow range. The instrument contained two very difficult items, six difficult items, four easy items, and three very easy items. This variation is desirable because it enables the instrument to measure students across different ability levels. Notably, however, the absence of items above 1.5 logits represents a coverage gap at the upper end of the difficulty continuum, limiting the instrument's capacity to precisely measure students with high critical thinking ability.

The predominance of difficult and very difficult items eight out of fifteen is consistent with the cognitively demanding nature of critical thinking tasks on renewable energy topics, which require students to evaluate complex trade-offs and draw evidence-based conclusions rather than retrieve factual information (Bao & Koenig, 2019). Nevertheless, future revisions should consider adding items at moderate difficulty levels (approximately -0.5 to 0.5 logits) to improve the spread of the difficulty distribution and enhance measurement precision across the mid-ability range.

Differential Item Functioning (DIF)

The DIF analysis demonstrated that all 15 items were free from statistically significant gender bias, with all Rasch-Welch probability values exceeding 0.05 , indicating that the instrument measured critical thinking skills equitably across male and female students. The complete results of the DIF analysis are presented in Table 9.

Table 9. Differential Item Functioning Analysis

Item	Logit (L)	Logit (P)	DIF Contrast	Prob. (t)	Interpretation
S1	0,40	0,07	0,33	0,582	No DIF
S2	0,62	1,01	-0,39	0,501	No DIF
S3	-0,59	0,07	-0,67	0,306	No DIF
S4	-0,07	0,68	-0,74	0,221	No DIF
S5	1,26	1,01	0,25	0,663	No DIF
S6	0,84	0,20	0,64	0,280	No DIF
S7	1,26	0,20	1,05	0,074	No DIF
S8	-1,60	-0,85	-0,76	0,332	No DIF

Item	Logit (L)	Logit (P)	DIF Contrast	Prob. (t)	Interpretation
S9	0,62	0,20	0,42	0,476	No DIF
S10	-0,07	-0,50	0,44	0,496	No DIF
S11	-0,89	0,33	-1,22	0,071	No DIF
S12	-0,89	-1,26	0,37	0,620	No DIF
S13	-0,06	0,20	-0,27	0,665	No DIF
S14	-0,59	-0,20	-0,39	0,550	No DIF
S15	-0,59	-1,26	0,67	0,357	No DIF

All fifteen items showed probability values exceeding 0.05, indicating that no item displayed statistically significant differential functioning between male and female students. This finding suggests that the instrument measured critical thinking skills equitably across gender groups, and that item performance was not systematically influenced by students' gender.

The absence of gender-based DIF in this instrument is meaningful for two reasons. First, it reflects that the item content grounded in renewable energy contexts such as solar panels, wind turbines, and energy policy did not introduce gender-specific advantages or disadvantages, a finding that aligns with studies suggesting that context-neutral or technology-oriented STEM items tend to minimize gender bias when constructed carefully (Ozdemir & Alshamrani, 2020; Susongko et al., 2021). Second, it provides empirical support for the instrument's fairness as an assessment tool in mixed-gender classroom settings, which is particularly relevant given the demographic composition of the present sample (26 male, 49 female students).

Although items S7 and S11 exhibited relatively large DIF contrasts of 1.05 and -1.22 logits respectively, neither reached statistical significance. The large contrast in S11 may warrant monitoring in future administrations with larger and more balanced gender samples, as the current sample's unequal gender ratio (approximately 1:2) may have reduced statistical power to detect subtle DIF effects (Penfield & Camilli 2006). Future studies with larger and more balanced gender samples are recommended to confirm the absence of DIF more conclusively.

Wright Map Analysis

The Wright Map analysis revealed a mistargeting condition in which the average student ability (1.05 logits) exceeded the average item difficulty (0.00 logits), indicating that the test items were generally easier relative to the overall ability of the student sample. Figure 1 presents the Wright Map of the critical thinking skills instrument on renewable energy, displaying the distribution of student ability and item difficulty on the same logit scale.

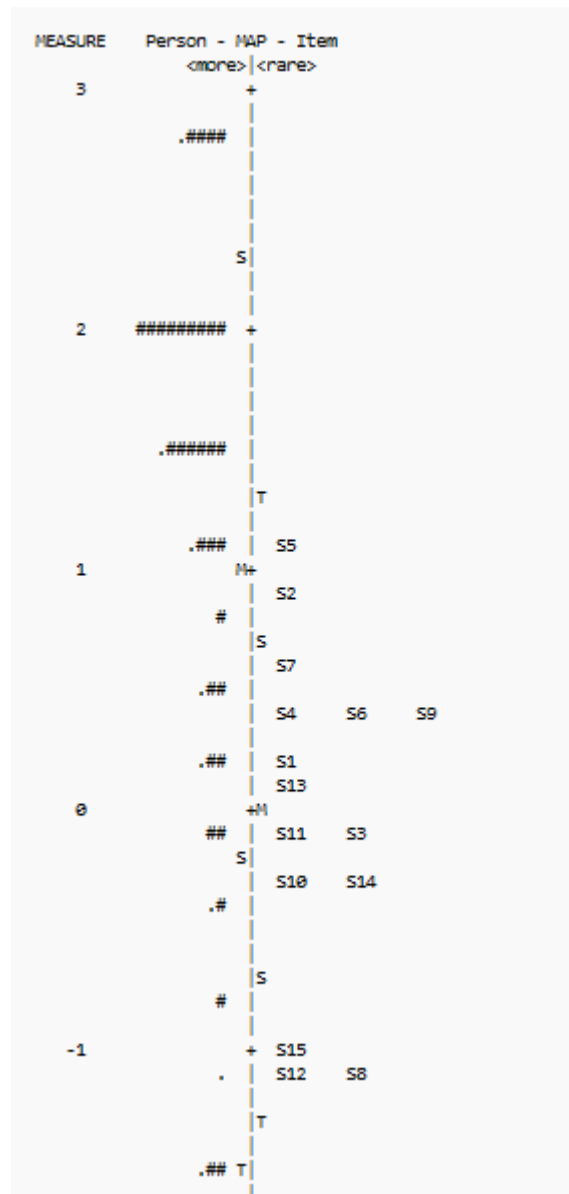


Figure 1. Wright Map of the Critical Thinking Skills Instrument on Renewable Energy

Based on Figure 1, the average student ability was 1.05 logits, higher than the average item difficulty set at 0.00 logits by Rasch model convention. This result indicates a mistargeting condition in which, on average, the test items were slightly easier relative to the overall ability of the student sample.

This finding requires careful interpretation in conjunction with the item difficulty distribution. While the average student ability exceeds the average item difficulty, this does not imply that all items were easy for all students. As shown in Table 8, eight of fifteen items were classified as difficult or very difficult (0.19 to 1.09 logits), and several students with ability estimates near or below the mean would have found these items genuinely challenging. The apparent contradiction is resolved by recognizing that the mean student ability (1.05 logits) exceeded even the most difficult item ($S_5 = 1.09$ logits) only marginally, and a non-trivial portion of students were distributed below the mean. Thus, the mistargeting is most pronounced at the upper tail of the ability distribution, where a cluster of high-ability students exceeded the ceiling of available items, rather than across the entire sample.

Item S_5 was identified as the most difficult item at 1.09 logits, while items S_8 and S_{12} were the easiest at -1.11 logits. The overall difficulty range of -1.11 to 1.09 logits spans approximately 2.2 logits, which is narrower than the recommended range of at least 4–5 logits for instruments intended to

differentiate students across a wide ability spectrum (Bond & Fox, 2007). This narrow range, combined with the concentration of student ability above the item ceiling, explains the moderate person reliability of 0.67 obtained in this study when items are insufficiently difficult relative to student ability, the model cannot separate high-ability students with adequate precision (Craft et al., 2020; Kassiavera et al., 2024).

To address these limitations, future revisions should incorporate additional high-difficulty items extending beyond the current item ceiling to better target students with advanced critical thinking ability. Item analysis showed that the highest-difficulty items were primarily represented within the Basic Clarification and Bases for a Decision dimensions, whereas the Advanced Clarification and Supposition and Integration dimensions were represented predominantly by easy to moderately difficult items. Consequently, these latter dimensions should be prioritized for the development of more challenging items requiring students to evaluate underlying assumptions, integrate multiple sources of scientific evidence, and justify complex decisions related to renewable energy. Expanding the difficulty range in these dimensions would improve measurement coverage at the upper end of the ability continuum, enhance person reliability, and strengthen the instrument's ability to differentiate students across a broader spectrum of critical thinking ability (Ennis, 2011).

Conclusion

This study provides a comprehensive empirical evaluation of a critical thinking skills instrument on renewable energy using the Rasch model. The results show that the instrument demonstrates acceptable unidimensionality, item fit, and reliability, with the majority of items functioning appropriately across ability levels and exhibiting gender fairness. Nonetheless, item S7 showed misfit, item S2 displayed weak discrimination, and the item set lacked sufficient coverage at the higher end of the ability continuum, limiting the instrument's precision in differentiating high-ability students.

Scientifically, this study contributes a more rigorous and multifaceted approach to critical thinking instrument evaluation by integrating several Rasch-based analyses, unidimensionality, item fit, reliability, item difficulty, Differential Item Functioning (DIF), and Wright Map interpretation, within the underexplored context of renewable energy in physics education. This integrated diagnostic framework extends beyond the conventional reliance on content validity and internal consistency alone, offering a replicable model for evaluating higher-order thinking instruments in science education more broadly.

These findings carry both practical and research implications. For educational practice, the instrument, once the identified items are revised, can support physics teachers and researchers in assessing students' critical thinking skills in classroom or intervention settings, particularly for formative or low-stakes purposes. For future research, further validation with larger and more diverse samples is recommended before the instrument is applied in high-stakes contexts, and the development of additional higher-difficulty items is needed to improve measurement precision for high-ability students. More broadly, this study underscores the value of adopting Rasch modeling as a standard quality-assurance procedure in the development of critical thinking and other higher-order thinking assessments.

References

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Alias, N., & Hafiz, M. (2021). Rasch model analysis on critical thinking test for secondary school students in Malaysia. *International Journal of Evaluation and Research in Education*, 10(3), 961–970. <https://doi.org/10.11591/ijere.v10i3.21312>
- Aradiansyah, R., Permanasari, A., Hernani, & Ramdhani, M. A. (2022). Development of critical thinking skills instrument in science using Rasch model. *Journal of Physics: Conference Series*, 2243(1).

- <https://doi.org/10.1088/1742-6596/2243/1/012072>
- Aviyanti, L., Saepuzaman, D., & Karim, S. (2023). Rasch analysis of a multiple-choice physics test on work and energy for senior high school students. *Jurnal Pendidikan Fisika Indonesia*, 19(1), 1–12. <https://doi.org/10.15294/jpfi.v19i1.35214>
- Bao, L., & Koenig, K. (2019). Physics education research for 21st century learning. *Disciplinary and Interdisciplinary Science Education Research*, 1(1). <https://doi.org/10.1186/s43031-019-0007-8>
- Bhakti, Y. B., Arthur, R., & Supriyati, Y. (2023). Development of an assessment instrument for critical thinking skills in Physics: A systematic literature review. In *Journal of Physics: Conference Series* (Vol. 2596, Number 1, pp. 1–8). <https://doi.org/10.1088/1742-6596/2596/1/012067>
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge. <https://doi.org/10.4324/9780429030499>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer.
- Care, E., Kim, H., Vista, A., & Anderson, K. (2018). *Education system alignment for 21st century skills: Focus on assessment*. Brookings Institution. https://www.brookings.edu/wp-content/uploads/2018/01/care-et-al_learning-metrics.pdf
- Craft, M. F., Feldon, D. F., & Brown, E. A. (2020). Instruction in critical thinking as a mechanism of transfer for adult learners. *Thinking Skills and Creativity*, 38. <https://doi.org/10.1016/j.tsc.2020.100726>
- Ennis, R. H. (2011). *The nature of critical thinking: An outline of critical thinking dispositions and abilities*. University of Illinois.
- Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction (The Delphi Report)*.
- Facione, P. A. (2020). *Critical thinking: What it is and why it counts*. Measured Reasons LLC. <https://www.insightassessment.com/content/download/1176/7580/file/CT-What&Why.pdf>
- Febriani, Jumadi, & Dwandaru, W. S. B. (2023). Socio-scientific issues in physics learning to improve students' critical thinking skills. *Revista Mexicana de Física E*, 20(1), 10202–10206. <https://doi.org/10.31349/RevMexFisE.20.010202>
- Halim, M. G., Saepuzaman, D., Aviyanti, L., Utama, J. A., & Nawas, A. (2026). Measuring critical thinking in physics: A Rasch analysis of instrument quality and gender equivalence. *Jurnal Pendidikan MIPA*, 27(1), 452–475. <https://doi.org/10.23960/jpmipa.v27i1.pp452-475>
- Kassiavera, S., Suparmi, A., Cari, C., & Sukarmin. (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>
- Kemendikbudristek. (2022). *Keputusan Kepala Badan Standar, Kurikulum, dan Asesmen Pendidikan Nomor 008/H/KR/2022 tentang Capaian Pembelajaran pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah pada Kurikulum Merdeka*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.
- Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions*, 7(4), 328. <https://www.rasch.org/rmt/rmt74b.htm>
- Linacre, J. M. (2016). *Winsteps® Rasch measurement computer program user's guide* (Version 3.). Winsteps.com. <https://www.winsteps.com>
- Mahir, M., & Martawijaya, M. A. (2025). The level of critical thinking skills of high school students in physics learning based on local wisdom. *Jurnal Edukasi Terkini*, 2(2), 26–33. <https://doi.org/10.70310/jet.2025.020206118>
- Maulana, I., Winarno, N., & Omar, M. (2020). Rasch model-based quality analysis of a critical thinking test instrument in optics concepts. *Journal of Science Learning*, 3(2), 111–119. <https://doi.org/10.17509/jsl.v3i2.19737>
- OECD. (2019). *PISA 2018 results (Volume I): What students know and can do*. OECD Publishing.

<https://doi.org/10.1787/5f07c754-en>

- Ozdemir, B., & Alshamrani, A. H. (2020). Examining the fairness of language test across gender with IRT-based differential item and test functioning methods. *International Journal of Learning, Teaching and Educational Research*, 19(6), 22–38. <https://doi.org/10.26803/ijlter.19.6.2>
- Paul, R., & Elder, L. (2008). *The miniature guide to critical thinking: Concepts and tools*. Foundation for Critical Thinking Press.
- Peabody, M. R., O'Neill, T. R., & Gregg, J. L. (2020). Effect of sample size on common item equating using the dichotomous Rasch model. *Applied Measurement in Education*, 33(1), 10–23. <https://doi.org/10.1080/08957347.2020.1712492>
- Penfield, R. D., & Camilli, G. (2006). Differential item functioning and item bias. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 125–167). Lawrence Erlbaum Associates.
- Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Parama Publishing.
- Sujatmika, S., Sutarno, Masykuri, M., & Prayitno, B. A. (2025). Applying the Rasch model to measure students' critical thinking skills on the science topic of the human circulatory system. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(4), em2622. <https://doi.org/10.29333/ejmste/16221>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Asesmen Pendidikan*. Trim Komunikata.
- Susongko, P., Arfiani, Y., & Kusuma, M. (2021). Determination of gender differential item functioning in Tegal students' scientific literacy skills with integrated science (SLiSIS) test using Rasch model. *Jurnal Pendidikan IPA Indonesia*, 10(2), 270–281. <https://doi.org/10.15294/jpii.v10i2.26775>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.
- Yuliarti, Y., Marlina, L., Siahaan, S. M., Fathurohman, A., & Sudirman, S. (2023). Profile of high school students' critical thinking skills about renewable energy materials. *Jurnal Penelitian Pendidikan IPA*, 9(11), 10151–10160. <https://doi.org/10.29303/jppipa.v9i11.5418>