Rasch Model Analysis of Creative Thinking Skills Test on Renewable Energy

Damsianus Sepulo Tukan[1], Parlindungan Sinaga[2*]
Lina Aviyanti[2], Duden Saepuzaman[2]

[1] *Study of Magister of Physics Education, Universitas Pendidikan Indonesia, Bandung 40254, Indonesia.*
[2] *Department of Physics Education, Universitas Pendidikan Indonesia, Bandung 40254, Indonesia.*

Email: psinaga@upi.edu*

Abstract

This study evaluated the psychometric properties of a Creative Thinking Skills (CTS) instrument designed for senior high school physics in the context of renewable energy using the Rasch Partial Credit Model (PCM). The instrument was constructed based on Torrance's four dimensions of creative thinking—fluency, flexibility, originality, and elaboration—and consisted of 15 essay items scored with a four-point analytic rubric. A psychometric validation study was conducted with 75 tenth-grade students from a senior high school in Bandung, Indonesia. Response data were analyzed using Winsteps version 4.8.2.0 to examine construct validity, item fit, item difficulty, reliability, unidimensionality, Wright Map targeting, and Differential Item Functioning (DIF). The findings demonstrated satisfactory psychometric properties, including a Cronbach's alpha (KR-20) of 0.87, person reliability of 0.89, item reliability of 0.98, and a raw variance explained by measures of 77.2%, indicating high internal consistency and unidimensionality. Most items met the Rasch model fit criteria and spanned an appropriate range of item difficulty. However, one item (S13) exhibited significant DIF and requires further qualitative review, while the Wright Map suggested that additional items at the extreme ends of the ability continuum could improve measurement precision. Overall, the findings indicate that the developed instrument provides reliable and valid measurement of students' creative thinking skills in renewable energy learning. Furthermore, the study highlights the usefulness of the Rasch PCM in evaluating item quality, measurement precision, and potential item bias. Future research should involve larger and more diverse samples and further investigate the DIF-flagged item. However, the findings should be interpreted cautiously because the validation was conducted using a limited sample from one school, and further studies with larger and more diverse samples are required.

Keywords: Rasch Partial Credit Model; Creative Thinking Skills; Instrument Validation; Physics Assessment; Renewable Energy.

INTRODUCTION

Renewable energy provides an appropriate context for assessing creative thinking because it integrates several core concepts in senior high school physics, including energy transformation, work and power, efficiency, electricity, heat transfer, and the environmental applications of physics. Tasks situated in this context require students to generate alternative technological solutions, compare different physical principles, evaluate efficiency and feasibility constraints, and elaborate scientifically reasoned designs. These cognitive demands correspond closely to the dimensions of fluency, flexibility, originality, and elaboration.

Despite its importance, previous studies have reported that students' creative thinking skills remain insufficiently developed. Students often experience difficulty generating multiple ideas, producing original solutions, and elaborating scientific explanations when solving physics problems (Batlolona et al., 2019; Leasa et al., 2021). Instructional approaches such as STEM education and Project-Based Learning have shown considerable potential to promote students' creative thinking skills (Sumarni & Kadarwati, 2020). However, the effectiveness of these approaches can only be interpreted accurately when researchers use assessment instruments with adequate psychometric quality.

Although previous studies have developed creative thinking instruments, most validation studies have relied on Classical Test Theory (CTT) or conventional validity procedures. Consequently, evidence regarding item functioning, person–item targeting, dimensionality, measurement precision, and Differential Item Functioning (DIF) remains limited, particularly for essay-based instruments in renewable energy learning. Therefore, a comprehensive psychometric validation using the Rasch Partial Credit Model is still needed.

The Rasch Partial Credit Model (PCM) was selected because it provides a comprehensive framework for evaluating essay-based instruments with polytomous scoring categories. In addition to estimating person ability and item difficulty on a common measurement scale, the PCM enables the evaluation of construct validity, item fit, reliability, unidimensionality, Wright Map targeting, response pattern consistency, and Differential Item Functioning (DIF). These analyses provide stronger psychometric evidence than conventional validation methods and support the development of fair and accurate educational assessments.

This study contributes to the literature by providing comprehensive psychometric validation of a Creative Thinking Skills (CTS) instrument using the Rasch Partial Credit Model, integrating construct validity, item fit, item difficulty hierarchy, Wright Map targeting, response pattern analysis, and Differential Item Functioning (DIF) into a single evaluation framework. This approach is expected to provide stronger evidence regarding the quality of creative thinking assessment in physics education and contribute to the development of more rigorous educational measurement practices.

Unlike previous studies that mainly reported reliability and validity indices using conventional approaches, this study provides a comprehensive Rasch-based evaluation by integrating item fit, item difficulty hierarchy, Wright Map targeting, residual-based unidimensionality, response pattern analysis, and Differential Item Functioning (DIF). This integrated approach provides stronger evidence regarding the quality and fairness of a creative thinking skills instrument in renewable-energy physics learning.

METHOD

Research Design

This study employed a quantitative psychometric validation design to evaluate the psychometric properties of a Creative Thinking Skills (CTS) instrument developed for renewable energy learning in senior high school physics. The instrument consisted of 15 essay items scored using a four-point analytic rubric. Response data were analyzed using the Rasch Partial Credit Model (PCM) implemented in Winsteps version 4.8.2.0. The PCM was selected because it is appropriate for analyzing polytomous response data and provides comprehensive evidence regarding item functioning and overall measurement quality, including construct validity, item fit, item difficulty calibration, reliability, unidimensionality, person–item targeting through the Wright Map, and Differential Item Functioning (DIF).

Participants and Instrument

The participants in this study were 75 tenth-grade students from two Grade 10 classes at a senior high school in Bandung, West Java, Indonesia. The participants were selected using convenience sampling, resulting in a sample of 34 male students (45.3%) and 41 female students (54.7%). Convenience sampling was employed because the study aimed to conduct a preliminary psychometric evaluation using an accessible group rather than to produce population-level estimates. A sample of 75 participants exceeded the commonly cited minimum of approximately 50 respondents for useful preliminary calibration of polytomous items under reasonably targeted conditions. However, because the participants came from one school and the person–item distribution showed imperfect targeting, the resulting item estimates were interpreted as preliminary and require cross-validation with larger and more heterogeneous samples (Linacre, 1994).

Participation in this study was voluntary. Students and their parents received information regarding the study objectives before data collection, and written informed consent was obtained. All responses were anonymized prior to analysis and were used exclusively for research purposes. Permission to conduct the study was granted by the school administration.

The initial instrument development was guided by Torrance's theoretical framework and by the alignment of each item with the intended creative thinking dimension. However, no formal external expert review or quantitative content-validity analysis was conducted before the empirical Rasch testing. Therefore, the present study provides theory-based and empirical psychometric evidence, but it does not claim formal expert-based content-validity evidence. Future instrument development should involve physics education and assessment experts to examine content relevance, representativeness, wording clarity, and the scientific accuracy of the renewable-energy contexts.

TABLE 1. Indicators of Creative Thinking Skills

Indicator of Creative Thinking Skills	Sub-indicator of Creative Thinking (Torrance, 1972)	Question Number
Fluency	The capacity to generate numerous relevant responses, ideas, or potential solutions when addressing a particular task or problem.	1, 5, 9, 15
Flexibility	The capacity to approach a problem from different perspectives and generate responses using varied categories, strategies, or lines of reasoning.	2, 6, 10, 13
Originality	The capacity to formulate uncommon, distinctive, or novel ideas that differ meaningfully from typical responses.	4, 8, 12
Elaboration	The capacity to expand an initial idea by adding relevant details, explanations, examples, and supporting information.	3, 7, 11, 14

The operational definitions were adapted and reformulated from Torrance's (1972) framework of creative thinking.

Scoring Rubric

Student responses were evaluated using a four-point analytic rubric adapted from Torrance's (1972) creative thinking framework. Each response was scored according to the extent to which it demonstrated the targeted creative thinking indicator. The rubric was applied consistently across all responses to ensure standardized scoring during the Rasch analysis. Higher scores reflected stronger evidence of creative thinking performance. The scoring criteria for each creative thinking dimension are presented in **TABLE 2.**

TABLE 2. Analytic Scoring Rubric for Creative Thinking Skills

Creative Thinking Dimension	Score 4	Score 3	Score 2	Score 1
Fluency	Generates a very large number of relevant ideas, concepts, or questions (≥20).	Generates many relevant ideas, concepts, or questions (15–19).	Generates a moderate number of relevant ideas, concepts, or questions (10–14).	Generates few relevant ideas, concepts, or questions (<10).
Flexibility	Produces highly diverse ideas representing multiple categories or perspectives.	Produces fairly diverse ideas representing several categories.	Produces limited diversity, with ideas belonging to similar categories.	Produces repetitive ideas with little or no variation.

Creative Thinking Dimension	Score 4	Score 3	Score 2	Score 1
Originality	Produces highly original, innovative, and uncommon ideas.	Produces reasonably original ideas with some novelty.	Produces ideas with limited originality and common characteristics.	Produces common ideas that closely resemble typical examples.
Elaboration	Provides highly detailed, systematic explanations, including clear descriptions of components and underlying physics principles.	Provides sufficiently detailed explanations with most important components included.	Provides limited detail, with several important components missing.	Provides very brief and unclear explanations with minimal detail.

Data Collection

The instrument was administered online using Google Forms to all participants under the same testing conditions. Participation in the study was voluntary, and all participants were informed about the purpose of the study before completing the instrument. To protect participants' privacy, responses were collected anonymously and used solely for research purposes. Students completed the instrument individually under standardized administration procedures. All responses were scored by one researcher using the analytic rubric presented in **TABLE 2**. No independent second scoring was conducted; therefore, inter-rater reliability could not be estimated. This limitation was considered when interpreting the psychometric results because the analysis did not separate item functioning from possible scorer-related variation. Future validation should employ at least two independent raters and report inter-rater agreement or rater-facet analysis.

Data Analysis

Before Rasch analysis, the dataset was screened for incomplete submissions, duplicate records, and scores outside the permitted range of 1 to 4. Before Rasch analysis, the dataset was screened for incomplete submissions, duplicate records, and scores outside the permitted range of 1 to 4. No incomplete submissions, duplicate records, or invalid scores were identified. Therefore, all 75 complete response patterns were retained for analysis.

The functioning of the rating categories was examined to ensure that the four scoring levels represented meaningful differences in students’ creative thinking performance. Category functioning was evaluated based on category frequency, average measures, outfit mean square statistics, and Andrich threshold ordering. A properly functioning rating scale should demonstrate increasing average measures across categories, acceptable category outfit statistics, and ordered thresholds. These analyses were conducted to confirm that the four-point analytic rubric provided appropriate measurement progression for the Rasch Partial Credit Model.

The response data were analyzed using the Rasch Partial Credit Model (PCM) implemented in Winsteps version 4.8.2.0. The analyses included summary statistics, person and item reliability, separation indices, item fit statistics, item difficulty calibration, person–item targeting using the Wright Map, unidimensionality assessment through Principal Component Analysis (PCA) of standardized residuals, and Differential Item Functioning (DIF). These analyses were conducted to evaluate the overall psychometric quality of the CTS instrument at both the person and item levels.

The functioning of the four scoring categories was evaluated using category frequencies, average person measures, category outfit MNSQ values, and Andrich thresholds. The categories were considered to function as intended when each category contained sufficient observations, the average measures increased monotonically from the lowest to the highest category, category outfit values did not indicate excessive noise, and the thresholds advanced in the expected order (Linacre, 2002).

The Rasch output was interpreted according to established Rasch measurement guidelines reported by Sumintono and Widhiarso (2015) and other recent Rasch studies. Item fit was evaluated using Outfit

Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point-Measure Correlation (PTMEA Corr.). Item fit was evaluated through a joint consideration of Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), Point-Measure Correlation (PTMEA Corr.), and conceptual alignment with the intended construct. Items with one or more statistics outside the recommended ranges were retained when the remaining indicators and theoretical considerations supported their measurement quality. Reliability was interpreted using person reliability, item reliability, separation indices, and Cronbach's alpha (KR-20). Item difficulty, unidimensionality, and Differential Item Functioning (DIF) were interpreted according to the criteria summarized in Tables 3–5.

TABLE 3. Reliability Interpretation Criteria

Index	Excellent	Good	Acceptable	Moderate	Poor
Person Reliability	>0.90	0.80–0.89	0.70–0.79	0.60–0.69	<0.60
Item Reliability	>0.90	0.80–0.89	0.70–0.79	0.60–0.69	<0.60
Cronbach's Alpha	≥0.90	0.80–0.89	0.70–0.79	—	<0.70

TABLE 4. Fit Statistics Interpretation Criteria

Statistic	Acceptable Range
Outfit MNSQ	0.5–1.5
Outfit ZSTD	-2.0 to +2.0
Point-Measure Correlation (PTMEA Corr.)	0.40–0.85

TABLE 5 summarizes the criteria used to evaluate unidimensionality based on the raw variance explained by the Rasch measures, the unexplained variance in the first contrast, and the eigenvalue of the first contrast.

TABLE 5. TABLE 5. Unidimensionality Interpretation Criteria

Analysis	Criterion	Interpretation
Raw variance explained by measures	≥20%	Acceptable
	≥40%	Good
	≥60%	Excellent
Unexplained variance in first contrast	<15%	Acceptable
First contrast eigenvalue	<3.0 eigenvalue units	Acceptable

RESULTS AND DISCUSSION

Result

The analyses included summary statistics and reliability, Wright Map analysis, item fit and item difficulty analysis, construct validity through Principal Component Analysis (PCA) of residuals, scalogram analysis, and Differential Item Functioning (DIF). Together, these analyses provide comprehensive evidence regarding the measurement quality, reliability, validity, and fairness of the developed instrument.

Summary Statistics and Reliability

The summary statistics of the Rasch analysis are presented in **TABLE 6**. The CTS instrument demonstrated good person reliability (0.89) and excellent item reliability (0.98), indicating consistent measurement of student ability and a highly stable hierarchy of item difficulty. The average person measure was −0.72 logits, whereas the average item difficulty was fixed at 0.00 logits by the Rasch calibration. This indicates that the respondents' average creative thinking ability was slightly lower than the overall difficulty of the instrument, suggesting that the instrument was moderately challenging for the participating students. The person separation index of 2.90 indicates that the instrument was capable of distinguishing respondents with different levels of creative thinking ability, whereas the item separation index of 6.85 confirmed a well-defined hierarchy of item difficulty. Furthermore, the Cronbach's alpha (KR-20) coefficient of 0.87 demonstrated good internal consistency between persons and items. The standard error of the person mean estimate was relatively small (0.17 logits), indicating stable estimation of participant ability under the Rasch model.

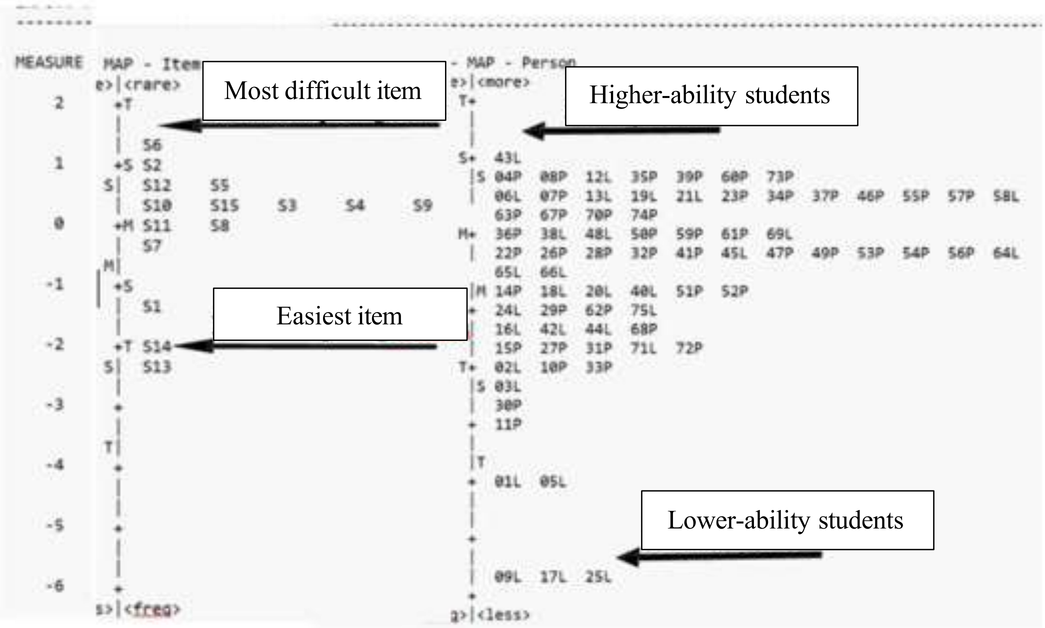
TABLE 6. Summary Statistics of the CTS Instrument

Parameter	Value	Interpretation
Number of participants	75	Grade 10 students
Number of items	15	Essay items
Person Reliability	0.89	Good
Item Reliability	0.98	Excellent
Person Separation	2.90	Good differentiation of student ability levels
Item Separation	6.85	Excellent hierarchy of item difficulty
Cronbach's Alpha (KR-20)	0.87	High internal consistency
Mean Person Measure (logit)	−0.72	Average person ability was slightly lower than the average item difficulty
Mean Item Measure (logit)	0.00	Reference point for Rasch item calibration

The relationship between person ability and item difficulty is further illustrated through the Wright Map analysis presented in **FIGURE 1.**

Wright Map Analysis

The Wright Map showed incomplete person–item targeting at both ends of the measurement continuum. Student ability ranged from approximately −5.5 to +1.5 logits, whereas item difficulty ranged from −2.50 to +1.38 logits. Item S13 was the easiest item, while S6 was the most difficult. Participants whose measures fell below −2.50 logits were not adequately targeted because no item represented their ability range. At the upper end, student ability extended slightly above the most difficult item, indicating limited measurement precision for the highest-ability participants. The targeting gap was substantially greater at the lower end. Therefore, additional items below −2.50 logits and several items above +1.38 logits should be developed to improve measurement precision across the full ability continuum.



Item Fit and Item Difficulty Analysis

The item difficulty estimates ranged from −2.50 to +1.38 logits. Item S6 was the most difficult item, whereas S13 was the easiest. Based on the joint interpretation of outfit MNSQ, outfit ZSTD, PTMEA correlation, item difficulty, and conceptual alignment, S1–S5, S7, S10, S13, and S14 generally met the stated item-fit criteria. Items S6, S8, S9, and S11 displayed overfit because their outfit MNSQ values were below 0.50 and their ZSTD values were below −2.00, indicating response patterns that were more predictable than expected by the model. Item S12 met the outfit MNSQ criterion but showed a slightly low ZSTD value of −2.12 and was therefore retained with a minor diagnostic flag. Item S15 displayed marginal underfit, with an outfit MNSQ value of 1.58, suggesting more unexplained response variation than expected. Item S13 fit the model statistically but was extremely easy and demonstrated significant DIF. Accordingly, all flagged items were retained only provisionally and require qualitative review and further empirical testing.

TABLE 7. Item Fit Statistics

Item	Measure (logit)	Outfit MNSQ	Outfit ZSTD	PTMEA Corr.	Decision
S1	-1.17	1.21	0.90	0.57	Fit
S2	0.91	1.07	0.35	0.43	Fit
S3	0.31	1.17	0.68	0.49	Fit
S4	0.44	1.36	1.28	0.48	Fit
S5	0.78	0.74	-0.97	0.54	Fit
S6	1.38	0.38	-2.59	0.55	Over Fit
S7	-0.28	0.58	-1.71	0.70	Fit
S8	0.10	0.48	-2.34	0.69	Slight overfit; retain and review
S9	0.35	0.46	-2.49	0.66	Overfit; retain and review
S10	0.31	0.60	-1.66	0.63	Fit
S11	0.10	0.48	-2.35	0.68	Slight overfit; retain and review
S12	0.75	0.51	-2.12	0.59	Minor ZSTD flag; retain and review
S13	-2.50	1.38	1.23	0.80	Fit; very easy; DIF review required
S14	-1.94	1.27	1.18	0.69	Fit; easy
S15	0.47	1.58	1.92	0.46	Marginal underfit; revise and retest

The decision was based on a joint evaluation of the fit statistics. Items with one or more statistics outside the recommended ranges were retained provisionally when their PTMEA correlation remained positive and conceptually consistent with the intended construct.

Construct Validity

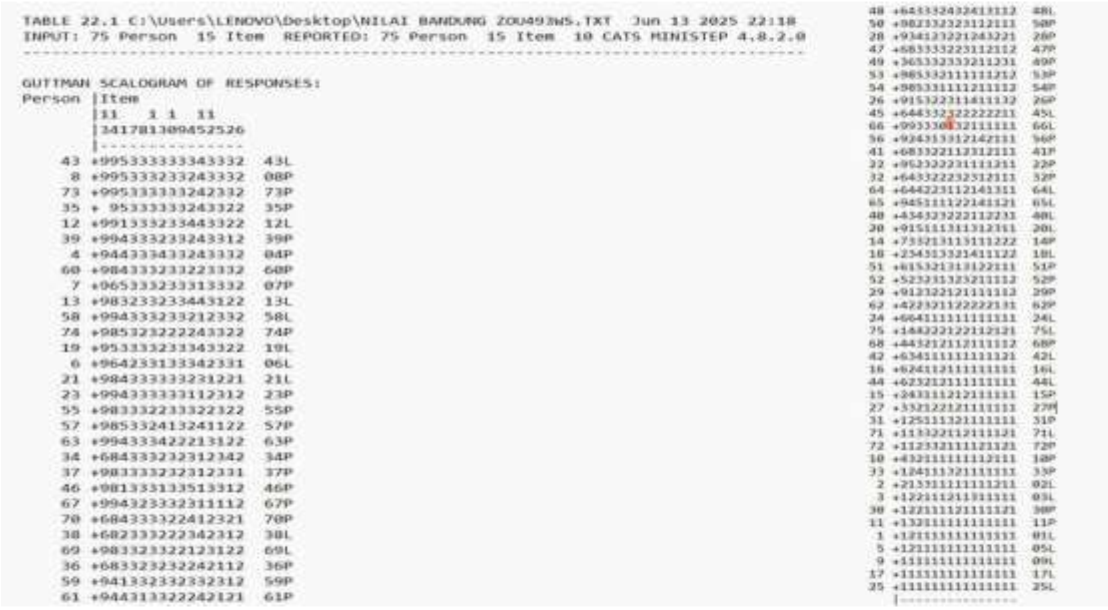
Items with overfit statistics were retained because their positive PTMEA correlations indicated alignment with the intended construct; however, overly predictable responses may indicate redundancy or scoring dependence. Item S15 requires review of its wording, scoring criteria, and construct alignment because its marginal underfit suggests unexplained response variation. Item S13 requires separate qualitative review because its significant DIF may reflect differences in wording, context familiarity, or cognitive demand between the two classes.

TABLE 8. Principal Component Analysis (PCA) of Rasch Residuals

Parameter	Value	Recommended Criterion	Interpretation
Raw Variance Explained by Measures	77.2%	>60%	Excellent
Unexplained Variance (1st Contrast)	3.1%	<15%	Excellent
Eigenvalue (1st Contrast)	2.07	<3.0	Accepted
Unexplained Variance (2nd Contrast)	2.8%	<15%	Accepted
Unexplained Variance (3rd Contrast)	2.5%	<15%	Accepted

Scalogram Analysis

The Guttman scalogram was examined as a descriptive representation of participants' response patterns. Participants with higher estimated abilities generally obtained higher scores across the items, whereas participants with lower estimated abilities generally obtained lower scores. However, a scalogram alone cannot establish person fit or confirm the absence of extreme person misfit. Person-level MNSQ and ZSTD statistics are required to support such a conclusion. Therefore, the scalogram was interpreted as complementary evidence of response-pattern consistency rather than as an independent person-fit test.



```
TABLE 22.1 C:\Users\LENOVO\Desktop\NILAI BANDUNG ZOU493WS.TXT  Jun 13 2025 22:18
INPUT: 75 Person  15 Item  REPORTED: 75 Person  15 Item  10 CATS MINISTEP 4.8.2.0
--------------------------------------------------------------------------------

GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |11   1 1  11
       |341781309452526
       |---------------
    43 +995333333343332  43L
     8 +995333233243332  08P
    73 +995333333242332  73P
    35 + 95333333243322  35P
    12 +991333233443322  12L
    39 +994333233243312  39P
     4 +944333433243332  04P
    60 +984333233223332  60P
     7 +965333233313332  07P
    13 +983233233443122  13L
    58 +994333233212332  58L
    74 +985323222243322  74P
    19 +953333233343322  19L
     6 +964233133342331  06L
    21 +984333333231221  21L
    23 +994333333112312  23P
    55 +983332233322322  55P
    57 +985332413241122  57P
    63 +994333422213122  63P
    34 +684333232312342  34P
    37 +983333232312331  37P
    46 +981333133513312  46P
    67 +994323332311112  67P
    70 +684333322412321  70P
    38 +682333222342312  38L
    69 +983323322123122  69L
    36 +683323232242112  36P
    59 +941332332332312  59P
    61 +944313322242121  61P
```

FIGURE 2. Guttman scalogram illustrating the response patterns of participants across the CTS instrument.

Differential Item Functioning (DIF)

Differential Item Functioning (DIF) analysis was conducted to examine whether items functioned differently between the two participating classes. Items with p-values below 0.05 were considered to exhibit significant DIF. The results showed that 14 of the 15 items exhibited no significant DIF ($p > 0.05$), indicating that these items functioned equivalently across groups. Only item S13 demonstrated significant DIF ($p = 0.002$), suggesting that students from the two classes responded differently to this item after controlling for overall creative thinking ability. Nevertheless, this finding does not necessarily indicate poor item quality but rather suggests that the wording or contextual characteristics of the item should be reviewed before the instrument is applied to broader populations. Overall, the DIF analysis indicates that the instrument demonstrated minimal evidence of item bias, although item S13 warrants further qualitative investigation.

TABLE 9. Differential Item Functioning (DIF)

Item	p-value	Interpretation
S1	0.981	No DIF
S2	0.493	No DIF
S3	0.894	No DIF
S4	0.303	No DIF
S5	0.437	No DIF
S6	0.967	No DIF
S7	0.778	No DIF
S8	0.671	No DIF

S9	0.722	No DIF
S10	0.550	No DIF
S11	0.714	No DIF
S12	0.824	No DIF
S13	0.002	Significant DIF
S14	0.444	No DIF
S15	0.295	No DIF

Items with $p < 0.05$ were considered to exhibit significant Differential Item Functioning (DIF). DIF significance was determined using the Rasch Welch t-test with a significance level of $p < 0.05$.

Discussion

The findings indicate that the Creative Thinking Skills (CTS) instrument demonstrates strong psychometric quality based on Rasch measurement. The instrument satisfied the major Rasch assumptions, including unidimensionality, reliable item calibration, appropriate person separation, and consistent differentiation of student ability. These findings are consistent with the principles of Rasch measurement proposed by Sumintono and Widhiarso (2015) and with recent Rasch-based psychometric validation studies (Juandi et al., 2024). Collectively, the results suggest that the instrument provides valid, reliable, and objective measures of creative thinking skills in renewable energy learning.

The Wright Map analysis indicated generally satisfactory targeting between item difficulty and participant ability. Although the average person measure (−0.72 logits) was lower than the average item difficulty (0.00 logits), most participants remained within the difficulty range represented by the instrument. However, several participants were located below the easiest item, suggesting that the addition of easier items could improve measurement precision for students with lower creative thinking ability. Overall, the instrument remained appropriately targeted for most participants.

The Differential Item Functioning (DIF) analysis provided additional evidence regarding instrument fairness. Fourteen items functioned equivalently across the two participating classes, whereas item S13 demonstrated statistically significant DIF. This finding suggests that the wording, context, or cognitive demands of S13 may have affected the two groups differently. Therefore, S13 should undergo qualitative review before the instrument is implemented in broader educational populations.

The item fit analysis generally supported the quality of the instrument. Most items demonstrated acceptable fit based on the joint evaluation of Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point-Measure Correlation (PTMEA Corr.). Several items showed one or more statistics outside the recommended ranges, while item S15 exhibited a marginally high Outfit MNSQ. Nevertheless, all items had positive PTMEA correlation values and remained conceptually aligned with the intended construct. Therefore, the flagged items were retained provisionally but should be reviewed qualitatively before broader use.

However, unlike most previous studies that primarily reported reliability and validity indices, the present study integrates Wright Map targeting, item fit statistics, item difficulty hierarchy, response pattern analysis, and DIF evaluation into a comprehensive psychometric assessment. This comprehensive approach provides robust psychometric evidence that the developed instrument is not only reliable and valid but also fair and suitable for assessing creative thinking skills in physics education, extending previous Rasch-based validation studies conducted by Juandi et al. (2024).

The present findings are more directly comparable with previous Rasch-based instrument studies than with instructional intervention research. Juandi et al. (2024) similarly applied Rasch analysis to examine the validity and reliability of a thinking-skills instrument in physics. Azizah and Wahyuningsih (2020) demonstrated the use of person and item fit statistics in evaluating a test instrument, while Rusmansyah and Almubarak (2020) showed how Rasch-based person and item information can support instructional planning. The present study extends this line of research by integrating item fit, item difficulty, Wright Map targeting, residual-based unidimensionality, scalogram inspection, and DIF analysis within one psychometric evaluation.

Beyond establishing the psychometric quality of the instrument, this study contributes to the development of creative thinking assessment in physics education by demonstrating the value of comprehensive Rasch-based validation. The integration of construct validity, item fit analysis, item difficulty hierarchy, Wright Map targeting, response pattern analysis, and Differential Item Functioning (DIF) evaluation provides a more rigorous framework for instrument development than conventional validation procedures. Consequently, the developed instrument can serve as a reliable assessment tool for both educational practice and future research on creative thinking skills in renewable energy learning.

From a practical perspective, the person and item measures can support differentiated instruction. Students whose measures fall below the easiest-item range should receive scaffolded tasks that emphasize generating several relevant ideas and explaining basic energy transformations before progressing to complex open-ended problems. Students near the center of the scale can be assigned tasks that require comparing multiple renewable-energy solutions and justifying their choices using efficiency, power, cost, and environmental constraints. Students whose measures approach or exceed the most difficult-item range require enrichment tasks that emphasize originality, optimization, and critical evaluation of competing designs. Teachers may also inspect student responses across the four creative-thinking indicators to identify specific areas requiring support. However, indicator-level interpretations should remain descriptive until the reliability and dimensionality of each indicator score have been evaluated separately.

Although minor refinements, including the addition of easier items and further qualitative review of the DIF-flagged item, may further improve measurement precision, the overall findings indicate that the instrument is psychometrically robust and suitable for broader implementation.

Several limitations should be acknowledged. First, the instrument was validated using responses from 75 Grade 10 students drawn from a single school, which may limit the generalizability of the findings. Second, although the Rasch analysis demonstrated strong psychometric performance, further validation involving more diverse samples and educational contexts is recommended.

In addition, all essay responses were scored by a single trained rater using a standardized analytic rubric. Although this procedure ensured scoring consistency throughout the study, future studies may involve multiple independent raters to examine inter-rater reliability and provide additional evidence for the consistency of the scoring process.

CONCLUSION

This study evaluated the psychometric properties of a Creative Thinking Skills instrument for renewable-energy learning using the Rasch Partial Credit Model. The results showed good person reliability, excellent item reliability, and strong evidence of unidimensionality. Most items functioned adequately according to the Rasch model. However, several items displayed overfit or marginal underfit, item S13 was extremely easy and demonstrated significant class-based DIF, and the Wright Map indicated insufficient coverage at the lower and upper ends of the ability continuum. Therefore, the results support the provisional use of the instrument in educational settings similar to the present research context, rather than unrestricted implementation across broader populations.

The hierarchy of item difficulty can assist teachers in designing progressive assessment tasks and differentiating instruction according to students' estimated ability levels. Nevertheless, interpretations should consider the limited sample from one school, the use of a single rater, the absence of formal expert content validation, the lack of inter-rater reliability evidence, and the limited number of items targeting students at the extreme ends of the ability continuum.

Future research should conduct expert content validation, evaluate the functioning of the four scoring categories, involve multiple independent raters, and cross-validate the item parameters using larger and more diverse samples. Further studies should also examine the quality of indicator-level scores and conduct expanded DIF analysis across gender, school type, geographic context, and socioeconomic background.

ACKNOWLEDGMENTS

The authors acknowledge the Indonesia Endowment Fund for Education (LPDP), Ministry of

Finance of the Republic of Indonesia, as the proposed funding institution for the presentation and dissemination of this research.

REFERENCES

Antonia, S.F.R. 2023, ‘A systematic review for creative thinking skills in physics subjects’, EduFisika: Jurnal Pendidikan Fisika, vol. 8, no. 2, pp. 154–163, doi: https://doi.org/10.59052/edufisika.v8i2.25281

Azizah, A. & Wahyuningsih, S. 2020, ‘Penggunaan model Rasch untuk analisis instrumen tes pada mata kuliah matematika aktuaria’, JUPITEK: Jurnal Pendidikan Matematika, vol. 3, no. 1, pp. 45–50, doi: https://doi.org/10.30598/jupitekvol3iss1pp45-50

Batlolona, J.R., Diantoro, M., Wartono & Latifah, E. 2019, ‘Creative thinking skills students in physics on solid material elasticity’, Journal of Turkish Science Education, vol. 16, no. 1, pp. 48–61, doi: 10.12973/tused.10265a.

Juandi, T., Kaniawati, I., Samsudin, A. & Riza, L.S. 2024, ‘The application of Rasch model to analyse the validity and reliability of an instrument for reflective thinking skills on topic of wave-particle dualism’, Kappa Journal, vol. 8, no. 2, pp. 270–277, doi: https://doi.org/10.29408/kpj.v8i2.27049

Linacre, J.M. 1994, ‘Sample size and item calibration [or person measure] stability’, Rasch Measurement Transactions, vol. 7, no. 4, p. 328.

Linacre, J.M. 2002, ‘Optimizing rating scale category effectiveness’, Journal of Applied Measurement, vol. 3, no. 1, pp. 85–106.

Leasa, M., Batlolona, J.R. & Talakua, M. 2021, ‘Elementary students’ creative thinking skills in science in the Maluku Islands, Indonesia’, Creativity Studies, vol. 14, no. 1, pp. 74–89. Doi: https://doi.org/10.3846/cs.2021.11244

Rusmansyah, R. & Almubarak, A. 2020, ‘Students’ cognitive analysis using Rasch modeling as an assessment for planning of strategies in chemistry learning’, JKPK: Jurnal Kimia dan Pendidikan Kimia, vol. 5, no. 3, pp. 222–235, doi: https://doi.org/10.20961/jkpk.v5i3.40936

Sumarni, W. & Kadarwati, S. 2020, ‘Ethno-STEM project-based learning: Its impact to critical and creative thinking skills’, Jurnal Pendidikan IPA Indonesia, vol. 9, no. 1, pp. 11–21. https://doi.org/10.15294/jpii.v9i1.21754

Sumintono, B. & Widhiarso, W. 2015, *Aplikasi pemodelan Rasch pada assessment pendidikan*, Trim Komunikata, Cimahi.

Torrance, E.P. 1972, ‘Predictive validity of the Torrance tests of creative thinking’, Journal of Creative Behavior, vol. 6, no. 4, pp. 236–262, doi: https://doi.org/10.1002/j.2162-6057.1972.tb00936.x