Kasuari: Physics Education Journal (KPEJ) 9(1) (2026) 186-202
P-ISSN: 2615-2681
E-ISSN: 2615-2673

Kasuari: Physics Education Journal (KPEJ)
University of Papua
website: https://journalfkipunipa.org/index.php/kpej

Development of an Ethnophysics-Based Critical Thinking Test on Heat Concept

Siti Nurhaliza*, Delthawati Isti Ratnaningtyas, I Komang Werdhiana, & Syamsuriwal
Physics Education, Faculty of Teacher Training and Education, Tadulako University, Indonesia
*Corresponding author: nurhalizaruslan891@gmail.com

Abstract: Critical thinking skills are important in physics education because they encourage students to analyze, evaluate, and apply concepts in real life. However, the assessment instruments used still focus on memory skills and have not integrated the local cultural context. In fact, many physical phenomena, including heat transfer, can be seen in cultural activities such as traditional food preparation. Previous studies have emphasized the development of questions based on higher-order thinking skills, literacy, and numeracy, but few have linked them to culture. Therefore, it is necessary to develop ethno-physics-based critical thinking questions on heat so that assessments not only measure conceptual mastery but also connect physics with local wisdom. This study aims to analyze the validity, reliability, discriminating power, and difficulty level of ethno-physics-based critical thinking questions on the topic of heat, as well as to test their practicality. The method applied is Tessmer's Research and Development (R&D) model, with stages including Preliminary, Self-Evaluation, and Formative Evaluation. The results of the development show that the resulting test items are valid. The reliability value obtained is 0.75, which is classified as high. In terms of difficulty, the majority of questions are categorized as moderate and a small portion are easy. The discrimination power of the test items varies between very good, good, and fair. Overall, all items successfully differentiated between students' abilities, and none of the items failed to differentiate. In terms of practicality, the use of these items is classified as very practical.

Keywords: critical thinking, ethnophysics, heat, Tessmer, test development

Pengembangan Soal Tes Kemampuan Berpikir Kritis Bermuatan Etnofisika pada Materi Kalor

Abstrak: Kemampuan berpikir kritis penting dalam pembelajaran fisika karena mendorong siswa menganalisis, mengevaluasi, dan menerapkan konsep dalam kehidupan nyata. Namun, instrumen penilaian yang digunakan masih berfokus pada kemampuan mengingat dan belum mengintegrasikan konteks budaya lokal. Padahal, banyak fenomena fisika, termasuk perpindahan kalor, terlihat dalam aktivitas budaya seperti pembuatan makanan tradisional. Penelitian sebelumnya lebih menekankan pada pengembangan soal berbasis HOTS, literasi, dan numerasi, tetapi belum banyak yang mengaitkannya dengan budaya. Oleh karena itu, perlu dikembangkan soal tes kemampuan berpikir kritis bermuatan etnofisika pada materi kalor, agar penilaian tidak hanya mengukur penguasaan konsep, tetapi juga menghubungkan fisika dengan kearifan lokal. Penelitian ini bertujuan untuk menganalisis validitas, reliabilitas, daya beda, dan tingkat kesulitan butir soal berpikir kritis berbasis etnofisika pada topik kalor, sekaligus menguji kepraktisan penggunaannya. Metode yang diterapkan adalah *Research and Development* (R&D) model Tessmer, dengan tahapan meliputi Preliminary, Self-Evaluation, dan Formative Evaluation. Hasil pengembangan menunjukkan bahwa instrumen soal yang dihasilkan dinyatakan valid. Nilai reliabilitas butir yang diperoleh sebesar 0,75 tergolong dalam kategori tinggi. Dari segi tingkat kesulitan, mayoritas soal berkategori sedang dan sebagian kecil termasuk mudah. Daya beda butir soal bervariasi antara kategori sangat baik, baik, dan cukup. Secara keseluruhan, semua butir soal berhasil membedakan kemampuan peserta didik, dan tidak ada satupun soal yang gagal dalam membedakan. Terkait aspek kepraktisan, penggunaan soal ini diklasifikasikan dalam kategori sangat praktis.

Kata kunci: berpikir kritis, etnofisika, kalor, pengembangan soal, Tessmer

INTRODUCTION

Critical thinking is a reasoning ability that demonstrates an individual's ability to evaluate phenomena scientifically and wisely from various perspectives to produce effective decisions Alberth et al., (2023). Critical thinking skills include the ability to make decisions based on logic and evidence by analyzing, evaluating, and synthesizing information. Critical thinking is very important in physics learning because this discipline requires a deep understanding of concepts and the application of scientific principles in solving real problems Amelia & Chusni, (2024). Physics is also known as a subject that can encourage the development of students' critical thinking skills (Syafitri et al., 2021).

The study of physics is fundamentally a cognitive process aimed at comprehending its concepts, principles, and laws, necessitating the cultivation of critical thinking skills in students (Tapanuli et al., 2018). These skills are crucial not only for grasping scientific ideas but also for connecting science to daily life, technology, and cultural contexts. Ultimately, critical thinking enables students to transfer academic knowledge into real-world applications (Risdianto et al., 2020).

Developed critical thinking skills need to be measured using appropriate assessment instruments to reflect students' actual abilities. A number of previous studies have developed instruments to measure high intelligence. The study conducted by Anggara and Ariawan (2022) produced questions based on critical thinking skills, but the development only reached the validity stage without field testing. Meanwhile, the study conducted by Kusuma & Nurmawanti (2023) developed literacy and numeracy questions based on high order thinking skills (HOTS) that were declared valid and in accordance with the principles of HOTS question development. Furthermore, Imamuddin et al., (2022) developed questions that were declared valid, practical, and effective. Based on the results of these studies, it can be concluded that the development of culture-based critical thinking questions is still limited. Previous studies have focused more on the development of literacy and numeracy questions, but not many have integrated the local cultural context into the assessment instruments. In fact, culture is an important aspect that develops along with technological advances, social interactions, and community adaptation to their environment. Culture encompasses seven key elements, namely language, science, society, technology, livelihood, religion, and art (Syakhrani & Kamil, 2022).

The connection between culture and science is evident in various community activities, one of which is the process of preparing traditional foods. For example, in making *ambal*, the dough is cooked using a clay pan called *dudongean* that stacks the dough in layers so that the heat is evenly distributed. This process involves heat transfer by conduction. Meanwhile, the making of *surabe* shows heat transfer through radiation, conduction, and convection simultaneously. This phenomenon shows a close relationship between physics concepts, especially heat, and local cultural practices.

Although several studies have attempted to develop culture-based learning and assessment instruments, most have been limited to strengthening literacy and numeracy aspects without specifically assessing students' critical thinking skills within the context of local culture. Furthermore, research integrating physics concepts especially heat with cultural elements in the form of critical thinking questions remains very limited. Therefore, it is necessary to conduct research focusing on the development of an ethnophysics-based critical thinking test instrument on heat material, so that the resulting assessment not only measures conceptual understanding but also reflects the interconnection between science and local cultural wisdom.

This study aims to test the validity, reliability, discriminating power, and difficulty level of critical thinking test questions with ethnophysics content on heat material and to

determine the practicality of using critical thinking test with ethnophysics content on heat material. The results of this study are expected to yield an assessment instrument that is valid, reliable, practical, and enriched with local wisdom, thereby deepening the meaning of physics learning and fostering students' appreciation of their cultural environment.

METHOD

The methodology for this study is Research and Development (R&D), implemented using the Tessmer model. This framework is structured around three core phases: preliminary study, self-evaluation, and formative evaluation. A systematic illustration of the entire research design is provided in Figure 1.

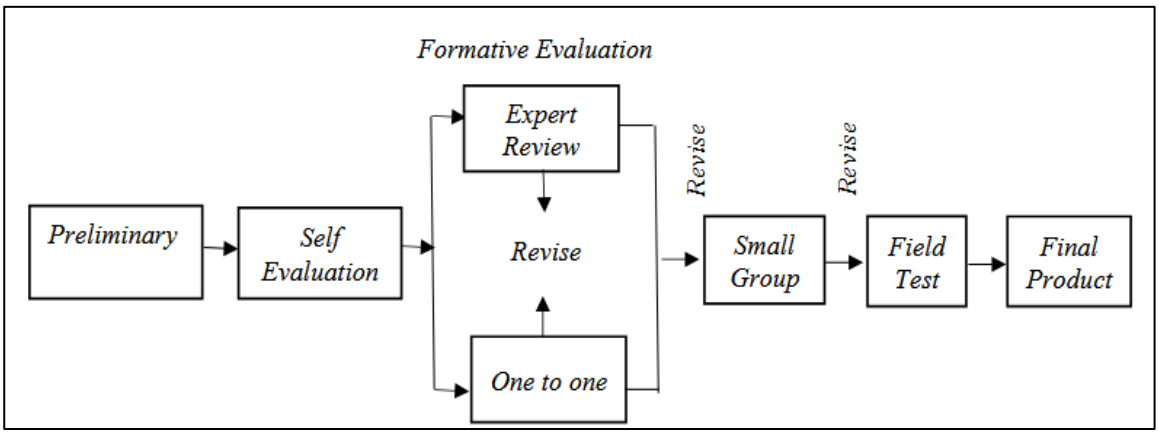

Figure 1. Research Design

The *Preliminary* stage was carried out by reviewing various reference sources related to assessment instruments, ethnophysics, and critical thinking, as well as conducting interviews and observations at schools to determine the assessment instruments used in physics learning. In the *Self-Evaluation* stage, the development of critical thinking test questions containing ethnophysics began with three stages of analysis, namely material analysis, curriculum analysis, and student characteristics analysis. Based on the results of these analyses, an outline was then compiled, followed by the formulation of questions and answer keys.

The Formative Evaluation stage involved testing three different groups. First, Expert Review or expert validation by two physics education lecturers and one physics teacher to assess the feasibility of the content, structure, and language. Next, individual tests (one-to-one) were conducted on three students to obtain direct feedback for revising the instrument design. The third stage was a small group test involving six students. Revisions from the previous two stages became the basis for improvements before this test was carried out, where the results and responses from students in this small group were then used to determine the final revisions to the product. This study proceeded to the Field Test phase, where the instrument was administered to 90 eleventh-grade students from SMAN Model Terpadu Madani Palu and SMAN 3 Palu. This stage aimed to empirically validate the instrument's validity, reliability, discriminating power, and difficulty level. The development process culminated in a Final Product comprising a set of valid, reliable, and practical ethno-physics-based critical thinking questions, established as an effective assessment tool for physics education.

The research was conducted in the even semester of the 2024/2025 academic year at two schools: SMA Negeri Model Terpadu Madani Palu and SMA Negeri 3 Palu.

Participants in the development stages included three validators for the expert review, three students for a one-to-one evaluation, and six more for a small group assessment. Additionally, a field test was administered to 90 eleventh-grade students from these schools, all of whom had already been taught the ethnophysics-integrated heat material.

The data analysis techniques were instrument validation, content and construct validation analysis, practicality data analysis, and empirical validity.

Instrument Validation

Instrument validation was carried out to ensure that the research instruments, namely the test questions, validation sheet, and student response questionnaire, were truly suitable for use in measuring what was to be studied using Equation 1.

$$\bar{X} = \frac{\Sigma x}{N} \quad (1)$$

Explanation:
$\bar{X}$ = Average score
Σx = Total assessment score for each aspect
N = Number of assessors

The scores obtained are used to determine the assessment criteria table with the provisions in Table 1.

Table1 . Instrument Validation Criteria

No	Interval	Score	Category
1	$X < \bar{X}_i + 1{,}8\ SB_i$	4.21≤ 5.00	Very Good
2	$\bar{X}_i + 0.6\ SB_i < X \leq \bar{X}_i + 1.8\ SB_i$	3.41 ≤ 4.20	Good
3	$\bar{X}_i - 0.6\ SB_i < X \leq + 0.6\ SB_i$	2.61 ≤ 3.40	Fair
4	$\bar{X}_i - 1.8\ SB_i < X \leq \bar{X}_i - 0.6\ SB_i$	1.81 ≤ 2.60	Less
5	$X \leq \bar{X}_i - 1.8\ SB_i$	X ≤ 1.80	Very Low

Description:
$\bar{X}_i$ = average ideal score
SB_i = ideal standard deviation
X = empirical score

Content and Construct Validity Analysis

Content validity analysis with Aiken's v coefficient using Equation 2. Test items are considered valid if they have a v value≥ 0.92.

$$V = \frac{\Sigma\ s}{[n\ (c-1)]} \quad (2)$$

Explanation:
V = Aiken validity index value
s = Score given by the assessor minus the lowest value on the assessment scale
r = The number given by the assessor
lo = Lowest score on the rating scale
n = Number of evaluators
c = Number of categories in the assessment

The reliability of the validators' assessment results is determined using Percentage Agreement with Equation 3, and the criteria can be seen in Table 2.

$$Percentage\ Agreement = \frac{Jumlah\ item\ kesepakatan\ penuh}{Total\ Item}\ X\ 100\ \% \quad (3)$$

Table 2. Percentage Agreement

No	Percentage Agreement Value (%)	Reliability Category
1	< 60	Poor
2	60–74	Moderate
3	75 – 89	Good
4	≥ 90	Very Good

Practicality data analysis

Students filled out a questionnaire regarding the practicality of the test questions. The method used to calculate practicality data was Equation 4, and the criteria for test question practicality can be seen in Table 3.

$$P = \frac{R}{SM} \times 100\% \quad (4)$$

Explanation:
P = Practicality score
R = Score obtained
SM = Maximum score

Table 3. Practicality Criteria for Questions (Wayan et.al, 2023)

No	Value	Criteria
1	$85 \leq P \leq 100$	Very Practical
2	$75 \leq P < 85$	Practical
3	$60 \leq P < 75$	Fairly Practical
4	$55 \leq P < 60$	Less Practical
5	$0 \leq P < 55$	Not Practical

Empirical Validity

The The evaluation of the instrument's empirical validity was performed via Rasch analysis, implemented through the Winsteps software. Model fit was gauged using the Outfit Mean Square (MNSQ) and Outfit Z-Standard (ZSTD) statistics. An item is classified as conforming to the Rasch model if it meets these two conditions:

1. Its Outfit MNSQ value is within the 0.5 to 1.5 interval.
2. Its Outfit ZSTD value falls between -2.0 and +2.0.

A reliable assessment instrument demonstrates consistency in its measurement outcomes. The reliability coefficient can be calculated using the Kuder-Richardson 20 (KR-20) formula. Interpretation of the results follows a standard classification: a coefficient exceeding 0.70 indicates satisfactory reliability. Alternatively, reliability can be determined through the Cronbach's Alpha method, as presented in Equation 5, with detailed categorization criteria provided in Table 4.

$$r = \frac{K}{(k-1)}\left(1 - \frac{\Sigma\sigma_b{}^2}{\sigma^t}\right) \quad (5)$$

Note:
r = Test reliability coefficient (Cronbach Alpha)
k = Number of questions
$\Sigma\sigma_b{}^2$ = Total item variance
σ^t = Total variance

Table 4. Item Reliability (Ramadhan et al., 2024)

No	Value	Category
1	$r_{11} < 0.20$	Very low
2	$0.20 \leq r_{11} < 0.40$	Low
3	$0.40 \leq r_{11} < 0.70$	Moderate
4	$0.70 \leq r_{11} < 0.90$	High
5	$0.90 \leq r_{11} < 1.00$	Very High

The item difficulty index (bi) generally ranges from -2.0 to +2.0 after the test takers' ability scores are standardized with a mean of 0 and a standard deviation of 1. An item can be categorized as very easy if the bi value is less than -2.0, and conversely, it can be categorized as very difficult if the bi value exceeds +2.0. More complete criteria regarding item difficulty levels are presented in Table 5.

Table 5. Item Difficulty Criteria (Rizbudiani et al., 2021)

No	Threshold Value	Description
1	$b > 2$	Very difficult
2	$1 < b \leq 2$	Difficult
3	$-1 \leq b \leq 1$	Moderate
4	$-2 \leq b < -1$	Easy
5	$b < -2$	Very Easy

Discrimination

Discrimination power serves as a metric for evaluating how effectively a test item differentiates between high-achieving and low-achieving students. This parameter reflects the item's relevance to the intended construct measurement. An irrelevant item suggests comprehension difficulties among respondents, necessitating its revision or removal. The appropriateness of an item can be assessed through the Point Measure Correlation (PTMEA CORR), whose interpretation follows the classification presented in Table 6.

Table 6. Discrimination Power (Rosli et al., 2020) ; Alagumalai et al., S 2005)

No	P Value	Category
1	≥ 0,40	Very good
2	0.30–0.39	Good
3	0.20 – 0.29	Fair
4	0.00 – 0.19	Failed to distinguish
5	< 0.00	Not applicable

RESULTS AND DISCUSSION

Results

This research successfully developed an ethnophysics-based critical thinking assessment on the concept of heat using the Tessmer model, which consists of three main phases: preliminary study, self-evaluation, and formative evaluation. The preliminary phase involved a comprehensive literature review, classroom observations, and interviews with physics teachers. The findings indicated that ethnophysics-based learning had already been implemented, highlighting the need for an assessment tool capable of measuring students' critical thinking skills within a local cultural context. These findings are consistent with the study conducted by Safitri et al., (2023) who developed a physics module based on local wisdom in the topic of temperature and heat, demonstrating that integrating cultural contexts into physics learning can enhance students' conceptual understanding and the relevance of the learning process. The assessment framework

developed in this study was grounded in Norris and Ennis's theoretical model of critical thinking, which encompasses five main dimensions: providing basic explanations, developing basic skills, making inferences, offering advanced explanations, and designing strategies and tactics Supriyati et al., (2018). This model is highly relevant to the ethnophysics context as it encourages students to analyze physical phenomena that emerge from daily life and local culture, enabling the development of critical thinking skills both cognitively and contextually in accordance with Indonesia's cultural diversity.

In the self-evaluation stage, the researchers conducted analysis and design activities. The analysis covered three main aspects, namely curriculum, students, and learning materials. The curriculum used in schools is the independent curriculum with a deep learning approach that emphasizes meaningful, critical, and contextual learning. The analysis of students indicates that 11th-grade learners are in the formal operational stage, a phase characterized by the development of abstract thinking and scientific reasoning skills. According to Piaget's theory, individuals aged approximately 13 to 17 begin to think logically and enter the formal operational stage marked by the emergence of abstract reasoning Ardiningtyas et al., (2022). The material analysis shows that the concept of heat is closely related to local cultural elements, such as the process of cooking *kaledo*, a dish typical of the Kaili tribe, which reflects the principle of heat transfer. Based on these analysis results, 30 multiple-choice questions with ethnophysics content were compiled with stimuli related to local culture and accompanied by images and narrative explanations.

The formative evaluation was carried out in a structured sequence of three stages: expert review, one-to-one evaluation, and small group assessment. A complete summary of the expert review's findings is presented in Table 7.

Table 7. Instrument Validation Results

Aspect	Average Score	Category
Material	4.50	Very Good
Construction	4.47	Very Good
Language	4.56	Very Good
Average Score	4.51	Very Good

Based on Table 7, the validation results indicate that the ethnophysics-based critical thinking assessment instrument on heat material, evaluated across three aspects content, construct, and language falls within the very good category, with the overall assessment also classified as very good.

The results of the content validity of the instrument, conducted using the Aiken V validity calculation technique, are presented in Table 8, which contains the validity coefficient values for each item.

Table 8. Aiken V Validity Results

Item	V	Description
Item 01	1.00	valid
Item 02	0.92	valid
Item 03	0.92	valid
Item 04	0.92	valid
Item 05	0.92	valid
Item 06	1.00	valid
Item 07	1.00	valid
Item 08	1.00	valid
Item 09	1.00	valid
Item 10	1.00	valid

Item	V	Description
Item 11	1.00	valid
Item 12	1.00	valid
Item 13	1.00	valid
Item 14	1.00	valid
Item 15	1.00	valid
Item 16	0.92	valid
Item 17	1.00	valid
Item 18	1.00	valid
Item 19	1.00	valid
Item 20	1.00	valid
Item 21	0.92	valid
Item 22	1.00	valid
Item 23	1.00	valid
Item 24	0.92	valid
Item 25	1.00	valid
Item 26	0.92	valid
Item 27	1.00	valid
Item 28	1.00	valid
Item 29	0.75	invalid
Item 30	0.75	Invalid

Based on Table 8, items 1 to 28 have an Aiken validity score between 0.92 and 1.00, indicating that all items are valid and suitable for use. Conversely, items 29 and 30 scored 0.75, thus failing to meet the validity criteria and declared invalid**.** During the field test stage, input from the small group test results formed the basis for product revision. The revised instrument was then tested on 90 11th grade science students at SMAN Model Terpadu Madani Palu and SMAN 3 Palu.

Item validity was examined through analysis of Outfit MNSQ and Outfit ZSTD values for each test question. The computational results derived from Rasch modeling are presented in Figure 2.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
19	25	90	1.46	.24	1.08	.8	1.12	1.0	.02	.21	72.2	72.2	E19
11	27	90	1.35	.24	1.11	1.1	1.14	1.2	-.03	.21	66.7	70.2	E11
6	39	90	.75	.22	.93	-1.3	.93	-1.3	.37	.22	67.8	60.9	E6
26	39	90	.75	.22	1.01	.1	1.01	.2	.20	.22	58.9	60.9	E26
9	40	90	.70	.22	1.05	.9	1.05	.8	.12	.22	55.6	60.4	E9
21	40	90	.70	.22	1.08	1.6	1.10	1.7	.03	.22	57.8	60.4	E21
28	40	90	.70	.22	.97	-.5	.97	-.6	.28	.22	62.2	60.4	E28
25	41	90	.65	.22	1.00	.1	1.00	.0	.21	.22	58.9	59.9	E25
16	47	90	.37	.22	1.00	.1	1.00	.0	.21	.22	58.9	58.9	E16
18	47	90	.37	.22	.89	-2.4	.88	-2.4	.46	.22	70.0	58.9	E18
14	49	90	.28	.22	.98	-.3	1.00	.0	.24	.22	64.4	59.3	E14
23	50	90	.23	.22	.93	-1.3	.91	-1.5	.37	.22	61.1	59.6	E23
2	51	90	.18	.22	.98	-.4	.98	-.2	.25	.22	62.2	59.8	E2
4	53	90	.09	.22	.97	-.6	.95	-.7	.29	.21	63.3	61.0	E4
13	54	90	.04	.22	1.02	.3	1.02	.3	.16	.21	57.8	61.6	E13
22	55	90	-.01	.22	1.06	1.0	1.10	1.3	.05	.21	61.1	62.4	E22
5	56	90	-.06	.22	.95	-.7	.93	-.9	.32	.21	62.2	63.2	E5
27	57	90	-.11	.22	.93	-1.0	.90	-1.2	.36	.21	67.8	64.0	E27
1	58	90	-.16	.22	.97	-.4	.95	-.5	.27	.21	66.7	64.9	E1
12	63	90	-.42	.23	.98	-.2	.99	.0	.21	.20	70.0	69.9	E12
24	63	90	-.42	.23	.99	-.1	.99	.0	.20	.20	70.0	69.9	E24
15	65	90	-.53	.24	1.04	.4	1.11	.9	.04	.19	72.2	72.1	E15
20	66	90	-.59	.24	1.11	.9	1.12	.9	.43	.19	68.9	73.1	E20
3	69	90	-.77	.25	.99	.0	.99	.0	.16	.18	76.7	76.4	E3
7	75	90	-1.19	.28	.99	.0	.99	.0	.12	.16	83.3	82.9	E7
17	75	90	-1.19	.28	.99	.0	1.05	.3	.09	.16	83.3	82.9	E17
8	76	90	-1.28	.29	1.00	.1	1.09	.5	.04	.16	84.4	84.0	E8
10	82	90	-1.88	.35	.92	-.2	.98	.1	.03	.13	91.1	90.2	E10
MEAN	53.6	90.0	.00	.24	1.00	-.1	1.01	.0			67.7	67.2	
S.D.	14.3	.0	.78	.03	.06	.8	.07	.9			8.9	8.9	

Figure 2. Validity Test Results

Analysis of Figure 2 reveals that most test items demonstrate acceptable fit with the Rasch model, fulfilling empirical validity criteria. The observed values MNSQ ranging from 0.88 to 1.14 and ZSTD between -2.4 and +1.7 largely conform to the acceptable thresholds of 0.5-1.5 for MNSQ and -2.0 to +2.0 for ZSTD. However, item number 18 was identified as an outlier with a ZSTD value of -2.4, rendering it invalid. Consequently, the final instrument comprises 27 items that exhibit satisfactory psychometric properties according to the Rasch measurement model.

The reliability of the instrument was assessed through a Rasch model analysis implemented in Winstep software, with the detailed output presented in Figure 3.

SUMMARY OF 28 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	53.6	90.0	.00	.24	1.00	-.1	1.01	.0
S.D.	14.3	.0	.78	.03	.06	.8	.07	.9
MAX.	82.0	90.0	1.46	.35	1.11	1.6	1.14	1.7
MIN.	25.0	90.0	-1.88	.22	.89	-2.4	.88	-2.4

REAL RMSE	.24	TRUE SD	.74	SEPARATION 3.09	ITEM RELIABILITY .91
MODEL RMSE	.24	TRUE SD	.74	SEPARATION 3.13	ITEM RELIABILITY .91
S.E. OF ITEM MEAN = .15					

Figure 3. Reliability Test Results

Based on Figure 3, the Winsteps analysis output shows an item reliability value of 0.91, which is classified as very high because it exceeds the minimum threshold of 0.70. This finding indicates that the instrument demonstrates strong internal consistency, making it suitable for consistently evaluating students' critical thinking skills. The high reliability value also suggests that the test items are capable of measuring the same construct in a stable manner, thereby supporting the instrument's internal validity. Theoretically, this aligns with Tessmer (1983) instrument development model, which emphasizes the importance of iterative evaluation to ensure that an instrument is not only content-valid but also internally reliable.

The analysis of item difficulty levels was performed utilizing Winsteps software, with the results graphically represented in the Wright Map presented in Figure 4.

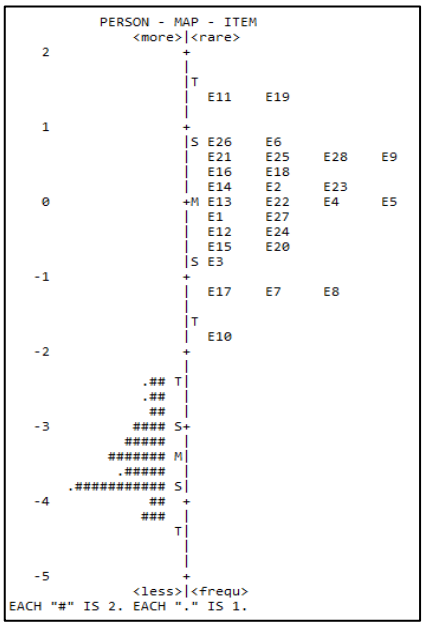

Figure 4. Difficulty Level

Based on the Wright Map, the distribution of items ranged from -2 to +2 logits. Items with logits > +1, such as E11 and E19, were classified as difficult, while items with logits < -1, such as E17, E7, E8, and E10, were classified as easy. Most items are in the range of -1 to +1, indicating that the majority of questions are of medium difficulty. Thus, the difficulty level of the questions has been arranged in proportion to the participants' abilities, and the variation in difficulty levels allows the instrument to effectively distinguish between low, medium, and high ability students.

The Point Measure Correlation (PTMEA CORR) value served as the basis for determining item discrimination in this study, the results of which are presented in Figure 5.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
19	25	90	1.46	.24	1.08	.8	1.12	1.0	.02	.21	72.2	72.2	E19
11	27	90	1.35	.24	1.11	1.1	1.14	1.2	-.03	.21	66.7	70.2	E11
6	39	90	.75	.22	.93	-1.3	.93	-1.3	.37	.22	67.8	60.9	E6
26	39	90	.75	.22	1.01	.1	1.01	.2	.20	.22	58.9	60.9	E26
9	40	90	.70	.22	1.05	.9	1.05	.8	.12	.22	55.6	60.4	E9
21	40	90	.70	.22	1.08	1.6	1.10	1.7	.03	.22	57.8	60.4	E21
28	40	90	.70	.22	.97	-.5	.97	-.6	.28	.22	62.2	60.4	E28
25	41	90	.65	.22	1.00	.1	1.00	.0	.21	.22	58.9	59.9	E25
16	47	90	.37	.22	1.00	.1	1.00	.0	.21	.22	58.9	58.9	E16
18	47	90	.37	.22	.89	-2.4	.88	-2.4	.46	.22	70.0	58.9	E18
14	49	90	.28	.22	.98	-.3	1.00	.0	.24	.22	64.4	59.3	E14
23	50	90	.23	.22	.93	-1.3	.91	-1.5	.37	.22	61.1	59.6	E23
2	51	90	.18	.22	.98	-.4	.98	-.2	.25	.22	62.2	59.8	E2
4	53	90	.09	.22	.97	-.6	.95	-.7	.29	.21	63.3	61.0	E4
13	54	90	.04	.22	1.02	.3	1.02	.3	.16	.21	57.8	61.6	E13
22	55	90	-.01	.22	1.06	1.0	1.10	1.3	.05	.21	61.1	62.4	E22
5	56	90	-.06	.22	.95	-.7	.93	-.9	.32	.21	62.2	63.2	E5
27	57	90	-.11	.22	.93	-1.0	.90	-1.2	.36	.21	67.8	64.0	E27
1	58	90	-.16	.22	.97	-.4	.95	-.5	.27	.21	66.7	64.9	E1
12	63	90	-.42	.23	.98	-.2	.99	.0	.21	.20	70.0	69.9	E12
24	63	90	-.42	.23	.99	-.1	.99	.0	.20	.20	70.0	69.9	E24
15	65	90	-.53	.24	1.04	.4	1.11	.9	.04	.19	72.2	72.1	E15
20	66	90	-.59	.24	1.11	.9	1.12	.9	.43	.19	68.9	73.1	E20
3	69	90	-.77	.25	.99	.0	.99	.0	.16	.18	76.7	76.4	E3
7	75	90	-1.19	.28	.99	.0	.99	.0	.12	.16	83.3	82.9	E7
17	75	90	-1.19	.28	.99	.0	1.05	.3	.09	.16	83.3	82.9	E17
8	76	90	-1.28	.29	1.00	.1	1.09	.5	.04	.16	84.4	84.0	E8
10	82	90	-1.88	.35	.92	-.2	.98	.1	.03	.13	91.1	90.2	E10
MEAN	53.6	90.0	.00	.24	1.00	-.1	1.01	.0			67.7	67.2	
S.D.	14.3	.0	.78	.03	.06	.8	.07	.9			8.9	8.9	

Figure 5. Discrimination Test Results

Based on Figure 5, the results of *the Point Measure Correlation* (PTMEA CORR) analysis show variations in the quality of discrimination between items. One item, E20, has a value of 0.43 and is classified as very good, while four items, E6, E23, E5, and E27, are in the range of 0.30-0.39 and are classified as good. The other ten items have values of 0.20–0.29 and are classified as adequate, so they are still suitable for use. Meanwhile, eleven items, namely E19, E9, E21, E13, E22, E15, E3, E7, E17, E8, and E10, with a PTMEA CORR value ≤ 0.19, as well as one item with a negative value of S11-0.03, were removed because they were unable to effectively differentiate students' abilities. Therefore, only questions with sufficient variation were retained to enable the instrument to accurately assess students' critical thinking abilities.

The field test yielded a finalized set of 15 ethnophysics-based critical thinking questions on heat material that satisfied all quality standards for validity, reliability, difficulty level, and discriminating power, as comprehensively detailed in Table 9.

Table 9. Results of Item Validity Testing

Item	Description
E1	valid
E2	valid
E3	invalid
E4	valid
E5	valid
E6	valid

Item	Description
E7	invalid
E8	invalid
E9	invalid
E10	Invalid
E11	invalid
E12	valid
E13	Invalid
E14	valid
E15	Invalid
E16	valid
E17	Invalid
E18	invalid
E19	Invalid
E20	valid
E21	Invalid
E22	invalid
E23	valid
E24	valid
E25	valid
E26	valid
E27	valid
E28	valid

Based on Table 9, of the 28 items tested, there were 13 invalid items: E3, E7, E8, E9, E10, E11, E13, E15, E17, E18, E19, E21, and E22. This invalidity was caused by items that were too easy (E8, E10, E15, E17) or too difficult (E11, E18, E19), so they were unable to optimally distinguish students' abilities. In addition, several items such as E3, E7, E9, E13, E21, and E22 had low PTMEA CORR correlations, which indicated that the answer patterns did not match the construct being measured. Based on the expert validation results, E3 and E17 were recommended for revision due to weaknesses in the answer choices that could potentially lead to multiple interpretations. However, due to fundamental invalidity, all invalid items were decided to be deleted so that the instrument would retain its good quality and strong validity. Furthermore, 15 valid items were retested to analyze their validity, reliability, difficulty level, and discrimination power.

Item validity analysis was conducted by examining the Outfit MNSQ and Outfit ZSTD values for each item using the Rasch model, as presented in Figure 6.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
5	39	90	.64	.23	.93	-.9	.89	-1.1	.44	.35	64.4	65.6	E5
13	39	90	.64	.23	1.09	1.1	1.14	1.4	.23	.35	62.2	65.6	E13
15	40	90	.59	.23	.94	-.7	.92	-.8	.42	.35	67.8	65.2	E15
12	41	90	.54	.23	1.10	1.3	1.11	1.1	.22	.35	60.0	64.9	E12
8	47	90	.24	.22	1.08	1.2	1.11	1.1	.24	.35	55.6	64.2	E8
7	49	90	.14	.22	1.01	.1	1.04	.4	.32	.35	64.4	64.2	E7
10	50	90	.08	.23	.89	-1.4	.84	-1.7	.47	.34	67.8	64.4	E10
2	51	90	.03	.23	.99	-.1	1.00	.0	.34	.34	64.4	64.7	E2
3	53	90	-.07	.23	.92	-1.0	.89	-1.0	.43	.34	71.1	65.1	E3
4	56	90	-.22	.23	1.00	.0	.95	-.4	.33	.34	66.7	66.4	E4
14	57	90	-.28	.23	.93	-.8	.92	-.6	.40	.33	67.8	67.0	E14
1	58	90	-.33	.23	1.10	1.1	1.07	.6	.21	.33	60.0	67.6	E1
6	63	90	-.61	.24	.95	-.4	.94	-.3	.34	.32	72.2	71.3	E6
11	63	90	-.61	.24	1.00	.1	1.07	.5	.26	.32	72.2	71.3	E11
9	66	90	-.79	.25	1.10	.9	1.33	1.8	.42	.31	71.1	73.9	E9
MEAN	51.5	90.0	.00	.23	1.00	.0	1.01	.1			65.9	66.8	
S.D.	8.8	.0	.46	.01	.07	.9	.12	1.0			4.8	2.9	

Figure 6. Final Product Validity Test Results

Based on Figure 6, all 15 items met the validity criteria based on Rasch model analysis using Winsteps software. The MNSQ values ranged from 0.84 to 1.33 and the ZSTD values ranged from -1.7 to +1.8, which are still within the ideal limits of MNSQ 0.5 - 1.5 and ZSTD -2.0 - +2.0. Thus, all items are empirically valid and consistent with the Rasch model. Reliability was analyzed using the Rasch model with Winsteps software, as shown in Figure 7.

SUMMARY OF 15 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	51.5	90.0	.00	.23	1.00	.0	1.01	.1
S.D.	8.8	.0	.46	.01	.07	.9	.12	1.0
MAX.	66.0	90.0	.64	.25	1.10	1.3	1.33	1.8
MIN.	39.0	90.0	-.79	.22	.89	-1.4	.84	-1.7

REAL RMSE .23 TRUE SD .39 SEPARATION 1.68 ITEM RELIABILITY .74
MODEL RMSE .23 TRUE SD .40 SEPARATION 1.72 ITEM RELIABILITY .75
S.E. OF ITEM MEAN = .12

Figure 7. Final Product Reliability Test Results

The output from Winsteps software presented in Figure 7 indicates a reliability coefficient of 0.75 for the test items, falling within the acceptable range for reliability. This result confirms the instrument's strong internal consistency, establishing its appropriateness for the consistent assessment of students' critical thinking abilities. The level of difficulty of the items was analyzed using Winsteps software by looking at the Wright map results, as shown in Figure 8.

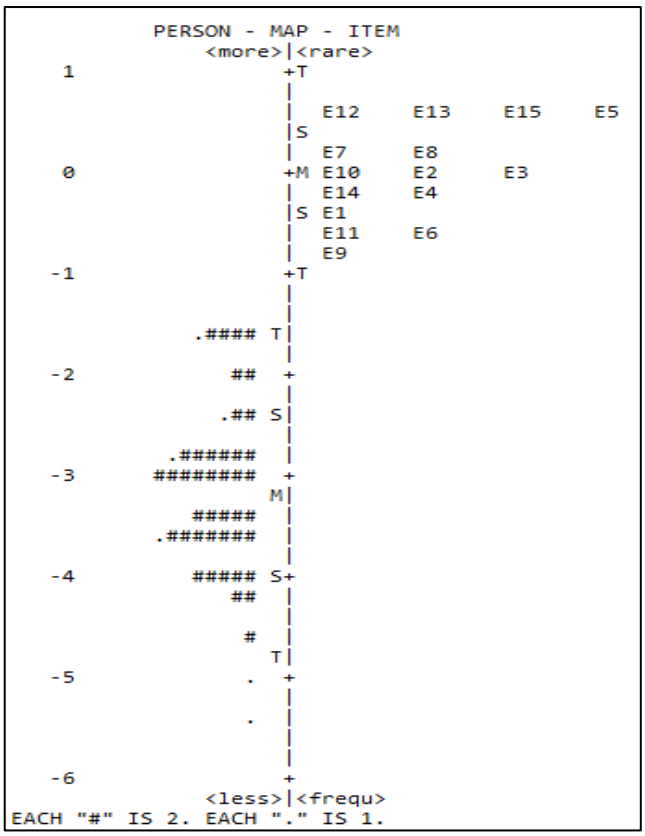

Figure 8. Final Product Difficulty Level Test Results

Based on Figure 8, the person item map reveals that all items are classified as moderate to easy in terms of difficulty. A total of 11 items are in the range of -1 < b < 1 (moderate category), while the other 4 items are in the interval of -2 < b < -1 (easy category). This finding indicates that the developed test instrument is proportional to the average ability level of students and is suitable for measuring critical thinking achievement at the intermediate level.

The analysis to determine the discriminating power of each item was performed using the Point Measure Correlation (PTMEA CORR), the results of which are displayed in Figure 9.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
5	39	90	.64	.23	.93	-.9	.89	-1.1	.44	.35	64.4	65.6	E5
13	39	90	.64	.23	1.09	1.1	1.14	1.4	.23	.35	62.2	65.6	E13
15	40	90	.59	.23	.94	-.7	.92	-.8	.42	.35	67.8	65.2	E15
12	41	90	.54	.23	1.10	1.3	1.11	1.1	.22	.35	60.0	64.9	E12
8	47	90	.24	.22	1.08	1.2	1.11	1.1	.24	.35	55.6	64.2	E8
7	49	90	.14	.22	1.01	.1	1.04	.4	.32	.35	64.4	64.2	E7
10	50	90	.08	.23	.89	-1.4	.84	-1.7	.47	.34	67.8	64.4	E10
2	51	90	.03	.23	.99	-.1	1.00	.0	.34	.34	64.4	64.7	E2
3	53	90	-.07	.23	.92	-1.0	.89	-1.0	.43	.34	71.1	65.1	E3
4	56	90	-.22	.23	1.00	.0	.95	-.4	.33	.34	66.7	66.4	E4
14	57	90	-.28	.23	.93	-.8	.92	-.6	.40	.33	67.8	67.0	E14
1	58	90	-.33	.23	1.10	1.1	1.07	.6	.21	.33	60.0	67.6	E1
6	63	90	-.61	.24	.95	-.4	.94	-.3	.34	.32	72.2	71.3	E6
11	63	90	-.61	.24	1.00	.1	1.07	.5	.26	.32	72.2	71.3	E11
9	66	90	-.79	.25	1.10	.9	1.33	1.8	.42	.31	71.1	73.9	E9
MEAN	51.5	90.0	.00	.23	1.00	.0	1.01	.1			65.9	66.8	
S.D.	8.8	.0	.46	.01	.07	.9	.12	1.0			4.8	2.9	

Figure 9. Final Product Discrimination Test Results

Based on Figure 9, the PTMEA CORR analysis results show that the items have varying discrimination, but items E5, E15, E10, E3, E14, and E9 are classified as very good with a correlation value above 0.40. four items, E7, E2, E4, E6, are categorized as good with a score of 0.30-0.39, and five items, E13, E12, E8, E1, E11, are categorized as sufficient with a score of 0.20-0.29. There are no items that fail to differentiate students' abilities, so overall, the instrument has good discrimination and is suitable for measuring critical thinking skills.

The practicality assessment of the instrument was conducted with 90 eleventh-grade students from SMA Model Terpadu Madani Palu and SMA Negeri 3 Palu. Quantitative analysis using Excel yielded a practicality score of 89.89%, which falls within the "very practical" classification.

Discussion

This research aims to evaluate the quality of a critical thinking assessment instrument designed for ethnophysics-integrated heat material, specifically examining its validity, reliability, difficulty level, discriminating power, and practicality. By implementing the Tessmer development model comprising preliminary study, self-evaluation, and formative evaluation stages the findings demonstrate that the developed instrument satisfies all required quality standards and functions effectively in measuring learners' critical thinking abilities.

Content validity analysis using the Aiken V index further strengthened the previous findings. A total of 28 items showed validity coefficients in the range of 0.92-1.00,

indicating substantial alignment with the target competencies. However, two other items, numbers 29 and 30, only achieved a value of 0.75 and were therefore declared invalid. Overall, these results prove that the majority of items have met the eligibility standards and can be implemented in instruments for measuring critical thinking skills in ethnophysics. Based on Rasch modeling, the majority of items showed compatibility with the measurement model used.

The Mean Square Outfit (MNSQ) values were recorded in the range of 0.84 to 1.33, while the ZSTD values were between -1.7 and +1.8. Both ranges are still within the acceptable tolerance limits, namely MNSQ 0.5-1.5 and ZSTD -2.0 to +2.0. Thus, the fifteen items that passed the selection were declared to meet empirical validity requirements.

The reliability analysis of the instrument produced an item coefficient of 0.75, which is classified as reliable, indicating the maintenance of the internal consistency of the instrument. This finding proves that each item functions stably and is able to measure the construct of students' critical thinking skills consistently. These results are consistent with the research Aprilia et al., (2023) hich reveals that evaluation tools with local cultural content can achieve a high level of reliability accompanied by adequate construct validity.

Based on the results of the Wright Map analysis, all test items fall within the moderate and easy categories. A total of 11 items E12, E13, E15, E5, E7, E8, E10, E2, E3, E14, and E4 are distributed within the range of $-1 < b < 1$, while 4 items E1, E11, E16, and E9 are within the range of $-2 < b < -1$. This distribution indicates that the developed items are proportional to the students' average ability and are effective in distinguishing different levels of performance. These findings are consistent with Handayani & Iba, (2020) who found that a science process skills test instrument with a balanced distribution of item difficulty consisting of six difficult, six moderate, and thirteen easy items was able to represent students' abilities proportionally.

The results of the PTMEA CORR analysis show that the discrimination power of the items is in the adequate category. There are 6 items, namely E5, E15, E10, E3, E14, and E9, which have very good discrimination power with $r > 0.40$. Meanwhile, 4 items, namely E7, E2, E4, and E6, fall into the good category with a range of 0.30–0.39, and 5 other items, namely E13, E12, E8, E1, and E11, are in the sufficient category with values between 0.20–0.29. No items with low discrimination power were found, indicating that all items are suitable for use. The varied composition of discrimination power shows that the developed instrument is capable of distinguishing students' abilities with a good level of effectiveness. This finding is consistent with the study of Hadmar et al, (2024) which showed that out of 20 test items for fifth-grade students at SD Negeri 2 Lamangga, 11 items had sufficient discrimination power (55%) and 9 items had good discrimination power (45%). These results indicate that a proportional variation in discrimination categories reflects a high-quality instrument that effectively differentiates students' levels of ability.

Based on the assessment conducted, the instrument obtained a practicality score of 89.89%, which is classified as very practical. This reflects that the material, presentation structure, and linguistic aspects of the questions can be well understood by students. These findings are in line with the results of the study Koza et al., (2024) which states that measurement tools based on local wisdom are able to elicit positive responses from students and achieve a level of practicality of up to 90%.

The findings of this study indicate that the critical thinking test items infused with ethnophysics on thermal phenomena are valid, reliable, practical, and proportional, making them a viable alternative for assessment in physics education. By integrating local cultural

practices, such as traditional cooking methods, the instrument not only assesses students' conceptual understanding but also promotes the development of higher-order thinking skills and scientific literacy in a contextualized manner. These results align with the studies of Aprilia et al., (2023) and Kurniahtunnisa et al., (2024), which emphasize the importance of incorporating local cultural elements into science assessment frameworks to enhance both scientific literacy and critical thinking skills.

Furthermore, Liwun et al., (2025) research demonstrates that an ethnoscience-based approach effectively bridges scientific knowledge with local cultural practices, such as the pottery-making process that involves concepts of heat and temperature. This supports the evidence that integrating local culture into physics learning not only improves students' conceptual understanding but also enriches their scientific literacy in real-life contexts. Consequently, the development of ethnophysics-infused critical thinking test items on thermal phenomena can serve as an effective strategy to improve the quality of physics education, strengthen students' scientific literacy, and simultaneously preserve local wisdom through education.

CONCLUSION AND SUGGESTIONS

Based on the research results, it can be concluded that the ethno-physics critical thinking test on heat material developed is valid and meets the eligibility criteria as a learning evaluation instrument. The reliability test results show an item reliability value of 0.75, which is in the high category, indicating that the questions have good internal consistency. As a result of the difficulty level analysis, most of the items were in the moderate category, while a small number were classified as easy, so that the questions were proportional in measuring students' critical thinking skills at an average level. In terms of discrimination power, all items were of good quality, with categories of very good, good, and sufficient, with no questions failing to distinguish students based on their ability level. In addition, the practicality test results show that the questions developed fall into the very practical category.

Based on the findings, it is recommended that educators and practitioners in the field of education implement this ethno-physics-based critical thinking test instrument as an evaluation option that not only measures conceptual understanding of physics but also stimulates students' critical thinking skills through the use of local wisdom that is contextual to their learning environment. For further research, it is recommended to develop similar tools for other subjects and educational levels, as well as to conduct trials with a broader and more diverse population to strengthen the external validity and generalization of the findings.

BIBLIOGRAPHY

Alagumalai, S., Curtis, D. D., & Hungi, N. (Eds.). (2005). *Applied Rasch Measurement: A Book of Exemplars: Papers in Honour of John P. Keeves* (Vol. 4). Springer. https://doi.org/10.1007/1-4020-3076-2

Alberth S. M., Fahrurrozi, E. U., & G. G. (2023). Implementasi berpikir kritis dalam upaya mengembangkan kemampuan berpikir kreatif mahasiswa. *Jurnal Papeda: Jurnal Publikasi Pendidikan Dasar, 5*(2)*,* 120-132. https://doi.org/10.36232/v5I2.3965

Amelia, N., & Chusni, M. M. (2024). Analisis Keterampilan Berpikir Kritis Dalam Pembelajaran Fisika Pada Materi Energi Terbarukan. *BIOCHEPHY: Journal of Science Education*, *4*(1), 248–252. https://doi.org/10.52562/biochephy.v4i1.1114

Anggara, R. P., & Ariawan, R. (2022). Pengembangan Soal Berbasis Kemampuan Berpikir Kritis Matematis Materi SPLTV Bernuansa Islami Kelas X. *Prisma*, *11*(1), 122-129.

https://doi.org/10.35194/jp.v11i1.1994

Aprilia, N., Setiani, Y., & Hadi FS, C. A. (2023). Pengembangan Instrumen Tes Numerasi pada Asesmen Kompetensi Minimum yang Bernilai Budaya Lokal. *Jurnal Educatio FKIP UNMA*, *9*(2), 850–857. https://doi.org/10.31949/educatio.v9i2.4824

Ardiningtyas, M., Harahap, T. H., & Panggabean, E. M. (2022). Penerapan Teori Piaget dalam Pembelajaran Matematika di Sekolah Menengah Atas *:* Studi Kasus di Sekolah SMA Negeri 3 Medan. *Tut Wuri Handayani: Jurnal Keguruan dan Ilmu Pendidikan*, *2*(2), 66–71. https://doi.org/10.59086/jkip.v2i2.294

Hadmar, S. S. A., Ali, A. M., & Yurfiah, Y. (2024). Analisis Daya Pembeda dan Tingkat Kesukaran Soal Pilihan Ganda pada Mata Pelajaran IPA di Sekolah Dasar. *PROSA: Jurnal Penelitian Pendidikan Guru Sekolah Dasar, 2*(3). https://doi.org/10.35326/prosa.v8i4.5368

Handayani, S. L., & Iba, K. (2020). Karakteristik Tes Keterampilan Proses Sains: Validitas, Reliabilitas, Tingkat Kesukaran dan Daya Pembeda Soal. *Publikasi Pendidikan*, *10*(2), 100-106. https://doi.org/10.26858/publikan.v10i2.13051

Imamuddin, M., Musril, H. A., & Isnaniah, I. (2022). Pengembangan Soal Literasi Matematika Terintegrasi Islam untuk Siswa Madrasah. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *11*(2), 1355-1371. https://doi.org/10.24127/ajpm.v11i2.4830

Koza, Y., Harso, A., & Doa, H. (2024). Pengembangan Instrumen Soal High Order Thinking Skill (HOTS) pada Materi Fluida Statis. *OPTIKA: Jurnal Pendidikan Fisika*, *8*(1), 69–78. https://doi.org/10.37478/optika.v8i1.3555

Kurniahtunnisa et al. (2024). Pengembangan Instrumen Tes Kemampuan Berpikir Kritis Materi Perubahan Iklim. *Eduproxima: Jurnal Ilmiah Pendidikan IPA*, *6*(2), 448–456. https://doi.org/10.29100/eduproxima.v6i2.5224

Kusuma, A. S., & Nurmawanti, I. (2023). Pengembangan Soal-Soal Literasi dan Numerasi Berbasis High Order Thinking Skills (HOTS) untuk Siswa Sekolah Dasar (SD). *Jurnal Ilmiah Profesi Pendidikan*, *8*(1), 516–523. https://doi.org/10.29303/jipp.v8i1.1313

Liwun, N. L., Huda, C., & Sayyadi, M. (2025). Universitas Papua Exploring Ethnoscience-Based Physics Concepts in the Pottery-Making Process of Kasongan Jogja: A Study on Heat and Temperature. *Kasuari: Physics Education Journal (KPEJ)* 8(1), 162–173. https://doi.org/10.37891/kpej.v8i1.936

Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas and Reliabilitas. *Journal on Education*, *6*(2), 10967–10975. https://doi.org/10.31004/joe.v6i2.4885

Rani, W. W., Caswita, S. S. (2023). Pengembangan LKPD Berbasis Inkuiri Terbimbing dengan Pendektan Kontekstual Berorientasi pada Kemampuan Pemecahan Masalah Matematis. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika 12*(2), 2327–2337. https://doi.org/10.24127/ajpm.v12i2.6734

Risdianto, E., Dinissjah, M. J., Nirwana, & Kristiawan, M. (2020). The effect of Ethno science-based direct instruction learning model in physics learning on students' critical thinking skill. *Universal Journal of Educational Research*, *8*(2), 611–615. https://doi.org/10.13189/ujer.2020.080233

Rizbudiani, A. D., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar: Jurnal Pendidikan Matematika, 12*(2), 399–412. https://doi.org/10.24042/ajpm.v12i2.9939

Rosli, R., Abdullah, M., Siregar, N. C., Abdul Hamid, N. S., Abdullah, S., Beng, G. K., Halim, L., Daud, N. M., Bahari, S. A., Majid, R. A., & Bais, B. (2020). Student

Awareness of Space Science: Rasch Model Analysis for Validity and Reliability. *World Journal of Education*, *10*(3), 170–180. https://doi.org/10.5430/wje.v10n3p170

Safitri, A. N., Sarwanto, S., & Harjunowibowo, D. (2023). Pengembangan Modul Pembelajaran Fisika Berbasis Kearifan Lokal Pada Materi Suhu dan Kalor. *Jurnal Materi dan Pembelajaran Fisika*, *13*(1), 32-39. https://doi.org/10.20961/jmpf.v13i1.60093

Supriyati, E., Ika Setyawati, O., Yuli Purwanti, D., Sirfa Salsabila, L., & Adi Prayitno, B. (2018). Profil Keterampilan Berpikir Kritis Siswa SMA Swasta di Sragen pada Materi Sistem Reproduksi Profile of Private High Schools Students' Critical Thinking Skills in Sragen on Reproductive System. *BIOEDUKASI: Jurnal Pendidikan Biologi*, *11*(2), 74–84. https://doi.org/10.20961/bioedukasi-uns.v11i2.21792

Syafitri, E., Armanto, D., & Rahmadani, E. (2021). Aksiologi Kemampuan Berpikir Kritis (Kajian tentang Manfaat dari Kemampuan Berpikir Kritis). *Journal of Science and Social Research*, *4*(3), 320-325. https://doi.org/10.54314/jssr.v4i3.682

Syakhrani, A. W., & Kamil, M. L. (2022). Budaya dan Kebudayaan: Tinjauan dari Berbagai Pakar, Wujud-Wujud Kebudayaan, 7 Unsur Kebudayaan yang Bersifat Universal. *Journal Form of Culture*, *5*(1), 1–10. https://journal.iaisambas.ac.id/index.php/Cross-Border/article/view/1161

Tapanuli, P., Hal, S., Wahyuni, S., Nasution, R., Pd, S., & Pd, M. (2018). Penerapan Model Inkuiri Terbimbing (Guided Inquiry) dalam Meningkatkan Kemampuan Berpikir Kritis pada Pembelajaran Fisika. *Jurnal Education and Development, 3*(1), 1–5. https://doi.org/10.37081/ed.v3i1.85

Tessmer, M. (1983). *Planning and Conducting Formative Evaluations*. Kogan Page.