Jurnal Penelitian Pembelajaran Fisika
Vol. 16 Issue 3 – July 2025, p310-324
p-ISSN 2086-2407, e-ISSN 2549-886X
Available Online at http://journal2.upgris.ac.id/index.php/JP2F
DOI: 10.26877/jp2f.v16i3.2112

Development of Holistic Assessment Instruments Based on Local Wisdom of Silver Crafts in Kota Gede on Temperature and Heat Topic

Rahmania Amanah Putri[1,2], Heru Kuswanto[1] and Edi Istiyono[1]

[1]Department of Physics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia

[2]E-mail: rahmania0013fmipa.2024@student.uny.ac.id

Received: 24 June 2025. Accepted: 25 July 2025. Published: 31 July 2025.

Abstract. This study aims to develop an assessment instrument based on local wisdom that covers three domains in physics learning: affective, cognitive, and psychomotor domains, especially on temperature and heat topic. This instrument was developed to assess students' understanding of scientific concepts and attitudes after receiving temperature and heat material using interactive e-books, and remains relevant to the demands of the Merdeka Curriculum. The research method uses the Research and Development (R&D) approach with development stages including planning, compilation, validation, and limited trials. Data collection techniques include tests and observations. The cognitive instrument is in the form of a pretest and posttest with multiple-choice item types. The affective instrument is a statement to assess scientific attitudes such as curiosity, critical thinking, creativity, openness to criticism, and cooperation. The psychomotor instrument is in the form of an observation sheet to assess students' presentation skills. The results of the validity and reliability tests show that all instruments are included in the good category and are suitable for use, but improvements are needed in terms of the number of respondents and the number of items to produce higher reliability values. The results of this assessment instrument development not only support instructional practices, but also its primary contribution lies in its ability to evaluate the three domains of student assessment comprehensively, while simultaneously enriching the scientific discourse on holistic assessment based on local wisdom in physics education.

Keywords: holistic assessment, local wisdom, temperature and heat

1. Introduction

Human resources are a key asset for national development [1]. The quality of education is a milestone in the progress of a nation. However, at the global level, the quality of education in Indonesia has yet to achieve satisfactory results. Indonesia still lags far behind developed countries in terms of the quality of its education [2]. The main challenge in advancing the quality of education centers on how education administrators implement innovation or reform. Innovation itself can be understood as a shift toward improvement. Therefore, to improve the quality of education, administrators must have a consistent drive to make continuous changes in line with developments in science and knowledge [3].

Popham argues that successful education is supported by accurate assessment, which has a positive impact on subsequent stages of learning. Assessment instruments, such as tests, must be developed with careful planning as a strategic step toward solving problems [4]. Choosing the appropriate assessment method will affect the evaluation results, making the evaluation process more objective and valid. Conversely, if the wrong assessment method is chosen, the information about the quality of education will be less accurate [5]. However, in practice, assessments conducted by teachers often focus only on cognitive aspects. Instruments for assessing attitudes and psychomotor skills have often not been developed systematically, resulting in assessments that are less valid and reliable. Therefore, efforts are needed to develop assessment instruments that are capable of measuring all three domains objectively and can be scientifically justified.

Education is more than just transferring information; it is also about absorbing cultural values from the surrounding environment. One of the main obstacles in the physics learning process is the ability of teachers to facilitate a deep understanding of concepts while instilling positive values related to local culture. Often, physics learning is not contextual, making it difficult for students to understand physics concepts correctly because they are not related to everyday life. Students can be said to have a good understanding of a concept if they can absorb the meaning of the material, reinterpret the concept in various situations, and apply it. Ethnophysics is a representation of the knowledge system that exists within a society, which includes values, activities, and cultural objects that are passed down from generation to generation [6]. Through ethno-physics-based learning, students' understanding of concepts will improve because they are trained to understand physics through the culture around them. Local wisdom-integrated testing instruments can connect physics material with local wisdom, enabling students to more easily understand the application of physics concepts in everyday life [7]. As part of the local cultural heritage, silver craftsmanship in Kotagede is a form of local wisdom that remains preserved and has cultural value rich in scientific principles, particularly related to temperature and heat. By using silver production as a learning context, students not only gain meaningful learning experiences, but are also introduced to the cultural values that exist in their region [8]. Therefore, it is necessary to develop comprehensive assessment instruments to measure students' conceptual understanding, scientific attitudes, and communication skills through learning that integrates local wisdom.

Several previous studies have been used as references for this study to develop holistic assessment instruments based on local wisdom. Related studies include research conducted by [9] about the development of fluid assessment using the Rasch model, which is part of modern IRT test theory. Another study conducted by [7] about the development of science literacy test instruments based on local wisdom. In addition, [2] also developed an assessment instrument for Higher Order Thinking Skills in physics lessons. Through previous research, the researchers finally found that although assessment instruments had been developed for various aspects, there was still room for innovation in research. Therefore, this study focused on developing a holistic, locally-based assessment of temperature and heat. This instrument is designed to measure three main domains in learning, namely: scientific attitude as the affective domain, conceptual understanding as the cognitive domain, and presentation skills as the psychomotor domain. The development of this instrument is expected to produce a valid, reliable, and contextual evaluation tool that can provide a comprehensive picture of student learning achievements in accordance with the holistic learning approach in the Merdeka Curriculum.

Based on the above description, this study is limited to the development of a holistic assessment instrument that covers three main domains, namely the cognitive, affective, and psychomotor domains in the subject of temperature and heat. The developed instrument focuses on the context of local wisdom in Kotagede silver crafts and is implemented through interactive e-book-based learning. The trial was conducted on a limited basis with 11th-grade students at SMA Negeri 1 Depok, so the results of this study are not intended for broad generalization but rather as a preliminary study that provides an overview of the quality of the developed instrument.

2. Method

The type of research used is Research and Development (R&D) with 4D development model modified into 3D. This modified 3D model consists of four stages, namely the definition stage (Define), the design stage (Design), and the limited development stage (Develop with Limitation) [10]. This model was chosen to suit the needs of developing affective, cognitive, and psychomotor assessment instruments to measure student achievement after participating in learning. This research was conducted in May 2025 at SMA Negeri 1 Depok with 33 students from class XI F4 who had already learned about temperature and heat as the research subjects. A one-group pretest-posttest Design is a research design that uses a single sample group and conducts measurements before and after treatment [11]. This design includes initial measurements (pretest) to measure the dependent variable before treatment is administered, followed by final measurements (posttest) to measure the effect of the treatment administered. Research subjects were selected using the Cluster Random Sampling technique. The stages in R&D research include the definition stage, the design stage, and limited validation and development.

The first stage is the design, which is carried out by interviewing physics teachers to identify the needs and challenges that require a solution. The second stage is the design of assessment instruments, which involves determining the assessment objectives and reviewing the literature to obtain operational definitions that include aspects, sub-aspects, and indicators. Once everything has been designed, the next step is to compile questions on conceptual understanding, scientific attitudes, and presentation skills. The assessment instrument to be developed is integrated with the local wisdom of Kotagede silver crafts in the subject of temperature and heat, which is adapted to the syntax of problem-based learning. This instrument is developed to measure students' scientific attitudes, conceptual understanding, and presentation skills after studying the physics of temperature and heat using interactive e-books. The third stage is product validation by experts. This stage involves two physics teacher practitioners who aim to assess, evaluate, and provide input on the instruments before conducting limited trials. The validity results can be determined through the scores obtained, which are calculated using the V Aiken formula.

$$V = \frac{\sum s}{n\,(c-1)} \tag{1}$$

The results of Aiken's quantitative calculations were then categorized qualitatively using Aiken's validity value interpretation. If the V value is < 0.4, the instrument's validity is low. Aiken's V value between 0.4 and 0.8 is considered to have moderate validity, while an Aiken's V value > 0.8 indicates high validity. Validity relates to a measure that indicates the level of validity of an instrument.

After passing the content validity test, the next step is to conduct a limited trial of the instrument involving 33 research subjects. The results of the limited instrument trial are then tested for validity, reliability, difficulty level, and discrimination using the item test in the Rasch model.

2.1. *Limited Test Instrument Validity*

The suitability criteria used were Infit/Outfit MNSQ values between 0.77 and 1.30, in accordance with the tolerance limits in the Rasch model. Validity with valid categories was seen in the INFIT MNSQ results with a range of 0.77 to 1.30, indicating that the instrument items were within the appropriate tolerance limits according to the Rasch model [12].

2.2. *Limited Test Instrument Reliability*

The results of the reliability analysis were examined using Rasch item analysis. Reliability values were reviewed based on item estimates and case estimates (respondents). Test instruments were categorized as having good reliability if they had reliability estimates greater than 0.70 [13].
The reliability values between respondents and items are presented in Table 1.

Table 1. Reliability value classification [13].

Value	Criteria
R ≥ 0,80	Very Reliable
0,60 < R ≤ 0,80	Reliable
0,40 < R ≤ 0,60	Fairly Reliable
0,20 < R ≤ 0,40	Less Reliable
R ≤ 0,20	Unreliable

2.3. *Item Difficulty and Discrimination Index Tests*

The results of the item difficulty analysis are seen in the threshold values in the item analysis. If the b value is greater than 2, the item is categorized as very difficult; conversely, if the b value is less than -2, the item is categorized as very easy. [14]. The level of difficulty can be seen from the threshold values presented in Table 2.

Table 2. Provisions on the level of difficulty of test items [14].

Value	Criteria
$b > 2$	Very difficult
$1 < b \leq 2$	Difficult
$-1 < b \leq 1$	Moderate
$-1 > b \geq 2$	Easy
$b < -2$	Very easy

The results of the item discrimination analysis were viewed using the iteman program. If the discrimination value ranges from 0.4 to 1.00 [15], then the item has good discriminating power. The criteria for item discriminating power are presented in Table 3.

Table 3. Point biserial correlation specifications [15].

Value	Criteria
0.4 – 1.00	Good
0.30 – 0.39	Moderate
0.20 – 0.29	Good enough (revision required)
Negatif -0.19	Poor (rejected)

In addition, after the assessment instrument has been declared valid and reliable, the effectiveness of the instrument in measuring students' understanding of concepts after studying physics material on temperature and heat using interactive e-books based on local wisdom can be tested. The analysis of product effectiveness in this study used repeated measures analysis. The research hypotheses included H0 and H1. H0 meant that there was no difference in conceptual understanding between the pretest and posttest scores. H1 meant that there was a difference in conceptual understanding between the pretest and posttest scores. Decisions on the hypotheses were based on the p-value. If the p-value was < 0.05, H0 was rejected. The magnitude of the difference in conceptual understanding between pretest and posttest scores is analyzed using effect size. The formula used in determining effect size is as follows :

$$effect\ size = \frac{Mean}{Standart\ Deviation} \tag{2}$$

Effect size categories are presented in Table 4 [16].

Table 4. Effect size range [16].

Effect Size (ES)	Category
$0.00 \leq \text{ES} < 0.20$	Ignored
$0.20 \leq \text{ES} < 0.50$	Small
$0.50 \leq \text{ES} < 0.80$	Moderate
$0.80 \leq \text{ES} < 1.30$	Large
$1.30 \leq \text{ES}$	Very Large

This study used 33 research subjects, which is relatively small, thus limiting the generalization of the results. However, this study is useful for preliminary research in developing a valid and reliable physics assessment instrument that is holistic in the affective, cognitive, and psychomotor domains.

3. Result and Discussion

The assessment instruments developed in this study included affective, cognitive, and psychomotor aspects. This assessment was given after students received material on temperature and heat, integrated with local wisdom. This assessment aimed to measure student achievement after receiving innovative learning on temperature and heat using e-books based on local wisdom. The instruments developed include three aspects of assessment. First, an attitude assessment instrument measures students' scientific attitudes in the form of a construct. Second, a cognitive assessment instrument with a construct measured in the form of students' conceptual understanding of the material studied. Third, a skill assessment instrument with a construct measured in the form of presentation skills. The instrument analysis technique uses a modern technique, namely the Rasch model of Item Response Theory (IRT). This study is in line with research [17] who developed instruments to measure higher-order thinking skills and scientific attitudes on the topic of reaction rates. In addition, other research conducted by [18] also developed a measurement instrument for higher-order thinking skills (HOTS) in elementary school students in Mongolia using Item Response Theory (IRT). The novelty of this research is the development of an instrument that serves as an evaluation tool in three aspects of holistic assessment (affective, cognitive, and psychomotor) integrated with local wisdom.

The researchers validated the instruments in terms of content before testing them on students. This validation included assessing the suitability of the content, construct, and language with the indicators being measured. Once the instruments had been validated in terms of content, the researchers conducted a limited trial involving 33 research subjects. The results of the content validity test on the three aspects of the instruments assessed by expert validators are presented as follows.

3.1. *Content Validity V Aiken*

The assessment instruments for scientific attitudes, conceptual understanding, and presentation skills have undergone content validity testing involving two physics teachers to assess the suitability of the product based on content, construct, and language. The testing method used was V Aiken. The test results are as follows:

3.1.1. Scientific attitude assessment instrument

Table 5. Scientific attitude assessment instrument validity test.

Validation Aspects	Number of indicators	Validator Score 1	Validator Score 2	Mean's V	Category
Content	2	8	6	0.75	Moderate
Construct	5	20	17	0.78	Moderate
Language	3	12	10	0.83	Very valid

The results of the V Aiken test on the scientific attitude assessment instrument were declared valid based on content, construct, and language aspects. The content validation aspect had a V Aiken value of 0.75, which is categorized as moderately valid. The construct aspect had a V Aiken value of 0.78, categorized as moderately valid. The language aspect had a V Aiken value of 0.83, categorized as highly valid. Thus, the interpretation of these findings is consistent with the research [19] which also used V Aiken and Rasch analysis in developing physics creativity instruments and obtained sufficient evidence of content validity. In addition, [20] In developing the locally-based instrument "Gasing," it was also reported that all items had a V Aiken value ≥ 0.78, indicating the experts' consistency regarding the relevance of the instrument items. Thus, the results of this study reinforce that the scientific attitude instrument developed has met the content validity criteria. Table 5 interprets that, overall, the scientific attitude affective instrument is valid in terms of content and is suitable for testing.

3.1.2. Concept understanding assessment instruments

Table 6. Concept understanding assessment instrument validity test.

Validity aspect	Number of indicators	Validator Score 1	Validator Score 2	Mean's V	Category
Content	3	149	147	0.983	Very valid
Construct	5	233	242	0.938	Very valid
Language	2	48	46	0.925	Very valid

The results of the V Aiken test on the concept comprehension assessment instrument were declared valid based on content, construct, and language aspects. Validators 1 and 2 scored high because each concept comprehension item was tested for content validity based on content, construct, and language. The content validation aspect had a V Aiken value of 0.983, which is categorized as very valid. The construct aspect had a V Aiken value of 0.938, which is categorized as very valid. The language aspect had a V Aiken value of 0.925, which is categorized as very valid. These findings exceed the minimum threshold that is generally used as a reference in research, as shown by the study [21] which obtained a V Aiken range of 0.64 to 0.93 on the analogical-transfer instrument for momentum and impulse, with all items declared valid. Thus, the results of this study provide evidence that the developed instrument has strong content validity, even higher than previous similar studies in the field of physics. Table 6 interprets that, overall, the concept understanding assessment instrument is content valid and feasible for testing.

3.1.3. Presentation assessment instruments

Next is the validity test of the presentation skill assessment instrument, as shown in Table 7.

Table 7. Communication assessment instrument validity test.

Validity aspect	Number of indicators	Validator Score 1	Validator Score 2	Mean's V	Category
Content	3	12	10	0.89	Very valid
Construct	3	12	10	0.89	Very valid
Language	3	12	8	0.78	Moderate

The results of the V Aiken test on the skill assessment instrument were declared valid based on content, construct, and language aspects. The content validation aspect had a V Aiken value of 0.89, categorized as very valid. The construct aspect had a V Aiken value of 0.89, categorized as very valid. The language aspect had a V Aiken value of 0.78, categorized as moderately valid. These findings are in line with previous research [22], which also reported a V Aiken value ≥ 0.78 on the physics communication skills instrument, both in verbal and nonverbal presentation aspects, so that the instrument was declared valid for use in learning. These results were also reinforced by [23], which found that the scientific argumentation instrument in physics learning had an average V Aiken value of 0.87, indicating a high level of content validity. Thus, the V Aiken value obtained in this study is consistent with previous studies and shows that the local wisdom-based presentation skill instrument has met the content validity criteria. Table 7 interprets that overall, the presentation skill assessment instrument is content valid and feasible to be tested. This confirms that the developed instrument is feasible to be used to assess students' presentation skills.

A comparison of the content validity test results in the three areas of the assessment instrument can be seen in Figure 1.

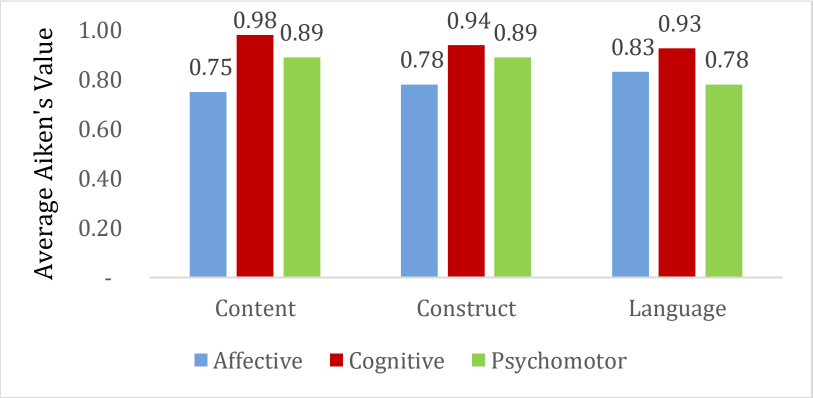

Figure 1. Comparison of V Aiken validity results in three domains.

The content validity results obtained using the V Aiken method show the comparison results for the three assessment domains shown in Figure 1. The highest content validity was found in the cognitive assessment instrument, with an average V Aiken score of 0.98. The highest construct validity was found in the cognitive instrument, with an average V Aiken score of 0.94. The highest language validity was found in the cognitive instrument, with an average V Aiken score of 0.93. According to [24] A test instrument is considered theoretically valid if it has a validity percentage of > 61%. Thus, all aspects of the assessment instrument are considered content valid.

3.2. *Analysis of affective Assessment instruments*

The analysis of affective assessment instruments includes testing the validity, reliability, difficulty level, and discrimination power of the instruments. The following is an explanation:

3.2.1. Item validity

Item validity analysis was conducted using the Rasch model, which produced Infit and Outfit Mean Square (MNSQ) values to evaluate the fit of each item to the model. The fit criterion used was an Infit/Outfit MNSQ value between 0.77 and 1.30. The results of the item validity analysis of the affective assessment instrument are presented in Table 8.

Table 8. Scientific attitude assessment instrument validity test.

Item	Infit MNSQ	Description	Decision
Item 1	1.02	Fit, Rasch model	Suitable for use
Item 2	1.00	Fit Ideal	Suitable for use
Item 3	1.02	Fit, Rasch model	Suitable for use
Item 4	1.17	Fit, Rasch model	Suitable for use
Item 5	0.81	Fit, Rasch model	Suitable for use
Item 6	1.00	Fit ideal	Suitable for use
Item 7	1.00	Fit ideal	Suitable for use
Item 8	0.98	Fit, Rasch model	Suitable for use
Item 9	1.04	Fit, Rasch model	Suitable for use
Item 10	0.94	Fit, Rasch model	Suitable for use

Table 8 shows that all attitude statement items are valid and consistent with the Rasch model. Based on the Mean Square Infit (MNSQ) analysis results shown in Table 8, all items in the attitude scale instrument have values that are still within the Rasch model tolerance limits, namely between 0.77 and 1.30. Items 2, 6, and 7 show an Infit value of 1.00, which is the ideal infit value, indicating that all three work very well in measuring the intended construct. Items 3, 4, 5, 8, 9, and 10 have Infit values in the range of 0.77–1.30. This fits the Rasch model [12]. This finding is reinforced by research [17] which states that items with an MNSQ infit value in the range of 0.77–1.33 are consistent with the construct being tested and are excellent in providing consistent results and information as expected. Overall, there were no misfit items, so all affective aspect items were considered valid and could be used further.

3.2.2. Item reliability

Furthermore, the results of the reliability test of the affective assessment instrument using the item test can be observed in Table 9.

Table 9. Scientific attitude instrument reliability score.

Mean	1.47
SD	0.98
SD (Adjusted)	0.80
Reliability of estimate	0.66

Table 9 shows the results of the affective instrument reliability test of 0.66. When viewed in the category range in Table 1, the instrument is considered reliable, but not yet meeting the criteria for highly reliable.

The low level of reliability is due to the insufficient number of respondents. If the sample is too small, the variety of answers given will also be limited. Thus, the items cannot yet cover individuals with high to low levels of ability, nor can they cover the range of difficulty from easy to difficult [9]. This instrument is a limited study as a form of preliminary research before developing instruments for broader research. These results are in line with the view that [13] which confirms that a reliability value in the range of 0.60 is still acceptable for exploratory studies. Therefore, although the reliability of this instrument is not yet optimal, its value is sufficient to support the use of the instrument for preliminary research, with the proviso that further development is needed to improve the consistency of the items in subsequent tests.

3.2.3. Level of difficulty and discrimination

The first test was the level of difficulty of the statement items. The level of difficulty of the items affected the probability of respondents giving answers. Table 10 provides information about the level of difficulty per item on the attitude statement instrument.

Table 10. Level of difficulty of scientific attitude items based on the Rasch model.

No	Thresholds (3-4)	Thresholds (4-5)	Level of Difficulty
1	-1.91	1.03	Moderate
2	-0.88	1.86	Difficult
3	-2.06	1.71	Moderate
4	-0.81	1.60	Difficult
5	-0.88	0.03	Moderate
6	-0.39	-0.19	Moderate
7	-0.75	-0.51	Moderate
8	-0.88	0.03	Moderate
9	-0.94	0.18	Moderate
10	0.13	0.28	Difficult

Threshold tests were observed from the frequency of answers selected by respondents. Based on the results of the affective instrument testing shown in Table 10, the answer categories selected by students were 3 and 4, so there were two threshold values, which were transitions from 3 to 4 and 4 to 5. Items 1, 3, 5, 6, 7, 8, and 9 have a moderate level of difficulty. Meanwhile, items 2, 4, and 10 have a high level of difficulty. The level of difficulty referred to in the attitude aspect is not the same as the level of difficulty in the cognitive assessment aspect. Each statement has a varying level of difficulty in describing the attitudes of students. Items with high logits require a stronger level of attitude to be agreed upon, while items with low logits tend to be easily agreed upon by the majority of students. The quality of a question can be seen from the level of difficulty of each question. A question is considered good if it is not too difficult and not too easy, or in other words, the level of difficulty of the question is moderate or sufficient [17]. [25] states that good quality questions are usually characterized by a distribution of 25% easy items, 50% medium items, and 25% difficult items. A similar view is expressed by [26], which emphasizes that the majority of items should be at a moderate level of difficulty so that the instrument can provide a more proportional picture of the students' abilities. Thus, the distribution of the difficulty level of the scientific attitude instrument items in this study can be said to meet the criteria for a good instrument. The following is the percentage of difficulty levels presented in Figure 2.

After examining the difficulty level of each item, the next step is to analyze the discrimination level of each item. Table 11 presents the results of the discrimination power analysis.

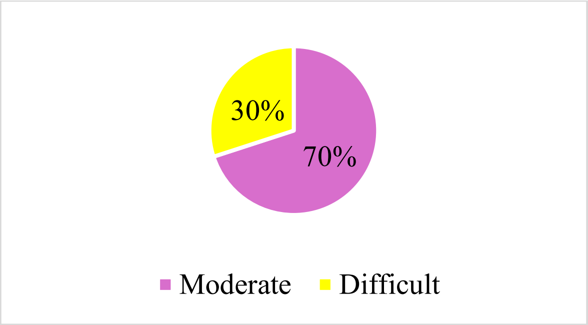

Figure 2. Level of difficulty of scientific attitude assessment instrument items.

Table 11. Analysis of item discrimination level.

No	Discrimination	Criteria
1	0.50	Good
2	0.48	Good
3	0.49	Good
4	0.55	Good
5	0.78	Very Good
6	0.87	Very Good
7	0.74	Very Good
8	0.69	Good
9	0.66	Good
10	0.69	Good

Based on the results of the analysis in Table 11, all items have a discrimination index ranging from 0.48 to 0.87. All items are categorized as good because they are above the minimum threshold of 0.40. Items 1, 4, 8, and 9 are in the range of 0.40–0.69, which means they have moderate discrimination but are still classified as good. Items 5, 6, 7, and 10 have a discrimination index > 0.70. This indicates excellent discriminatory ability for students with different ability levels. [25] states that items with a discrimination index of 0.41–0.70 are classified as good and those with a discrimination index of ≥0.71 are classified as excellent. This is in line with the opinion of [26] which states that items with a discrimination index ≥0.40 are of good quality. The overall interpretation states that these items are of good quality and suitable for measuring scientific attitude constructs. The following is the percentage of the item discrimination level presented in Figure 3.

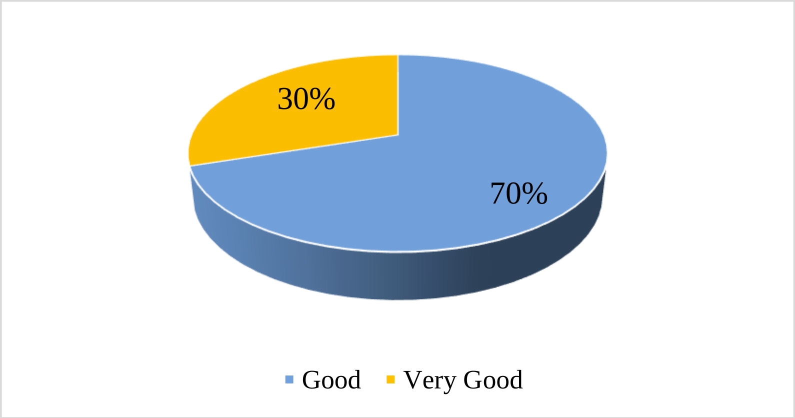

Figure 3. Analysis of item discrimination level.

3.3. *Analysis of Cognitive Assessment Instruments*

3.3.1. Item validity

Item validity analysis was conducted using the Rasch model, which produced Infit and Outfit Mean Square (MNSQ) values to evaluate the fit of each item to the model. The fit criterion used was an Infit/Outfit MNSQ value between 0.77 and 1.30. The results of the item validity analysis of the cognitive assessment instrument for concept understanding are presented in Table 12.

Table 12. Concept understanding assessment instrument validity test.

Item	Infit MNSQ	Description	Decision
Item 1	1.05	Fit, Rasch Model	Suitable for use
Item 2	0.89	Fit, Rasch Model	Suitable for use
Item 3	1.03	Fit, Rasch Model	Suitable for use
Item 4	1.03	Fit, Rasch Model	Suitable for use
Item 5	0.92	Fit, Rasch Model	Suitable for use
Item 6	1.01	Fit ideal	Suitable for use
Item 7	0.88	Fit, Rasch Model	Suitable for use
Item 8	1.05	Fit, Rasch Model	Suitable for use
Item 9	0.96	Fit, Rasch Model	Suitable for use
Item 10	1.17	Fit, Rasch Model	Suitable for use

Table 12 shows that each item is valid and consistent with the Rasch model. Based on the Mean Square Infit (MNSQ) analysis results shown in Table 12, all items in the attitude scale instrument have values that are still within the Rasch model tolerance limits, namely between 0.77 and 1.30. Items 1, 3, 4, 8, and 10 show Infit values above 1.00. Items 2, 5, 7, and 9 show Infit values below 1 and are still within the range of fit with the Rasch model. Item number 6 has an MNSQ Infit value of 1.01, indicating that item number 6 is highly fit with the Rasch model. All items are fit within the range of the Rasch model [12]. This finding is reinforced by research [17], which states that items with an MNSQ infit value in the range of 0.77–1.33 are consistent with the construct being tested and are excellent at providing consistent results and information as expected. Overall, there were no misfit items, so all cognitive aspect items were considered valid and could be used further.

3.3.2. Item reliability

The results of the reliability test of the cognitive assessment instrument using the item test can be seen in Table 13.

Table 13. Concept understanding instrument reliability score.

Mean	1.41
SD	0.65
SD (Adjusted)	0.48
Reliability of estimate	0.51

The reliability value of the concept understanding test items tested using Iteman is 0.51. When viewed in the reliability criteria range in Table 1, the reliability value of the concept understanding items is classified as less reliable because it is less than 0.6. This value is still below the minimum standard generally used, which is ≥0.70 to indicate the internal consistency of the instrument [27]. Based on reliability interpretation criteria, values <0.50 are categorized as unacceptable, values between 0.60 and 0.70 are still acceptable in a limited context, and values ≥0.80 are considered high [28]. The low reliability score in this study was influenced by the relatively small number of respondents (33 students), resulting in limited response variation and suboptimal internal consistency [29]. This was also conveyed in research conducted by [30] If the sample size is too small, the variety of responses will also be limited.

Therefore, a large number of research subjects is needed to determine the reliability of an assessment instrument in further research. This existing concept understanding instrument can still be used in limited tests with a small number of respondents. Further review is needed if a large-scale test is to be conducted.

3.3.3. Level of difficulty and discrimination index of test items

This test was conducted using an item test on the Rasch model so that the level of difficulty and discrimination of each item could be determined. The results of the item difficulty level analysis are presented in Table 14. This test was conducted to determine the students' abilities based on the probability of students answering the items.

Table 14. Level of difficulty of conceptual understanding items.

No	Thresholds	Level of Difficulty
1	-0.36	Easy
2	-0.19	Easy
3	0.11	Difficult
5	0.11	Difficult
6	-0.19	Easy
7	1.08	Very Difficult
8	-0.74	Easy
9	0.11	Difficult
10	0.81	Difficult

The results of the analysis in Table 14 show that the level of difficulty of the items varies considerably. Of the nine items analyzed, four are classified as easy, namely items 1, 2, 6, and 8. The items classified as difficult are items 3, 5, 9, and 10. Overall, these items are suitable for measuring the construct of concept understanding in students. If the questions are too easy, the students' abilities cannot be determined because these items are not challenging enough and will not stimulate the students' development in terms of their efforts to solve the problems. Meanwhile, questions that are too difficult will make students feel discouraged and unmotivated in finding solutions to the problems they face [31]. The quality of questions can be seen from the level of difficulty of each question. A question is considered good if it is not too difficult and not too easy, or in other words, if the level of difficulty is moderate or sufficient [17]. Therefore, item difficulty analysis is also necessary to ensure that the items provided are appropriate for the students' abilities and able to comprehensively represent their understanding of the concepts. Figure 4 shows the percentage of item difficulty levels in concept understanding.

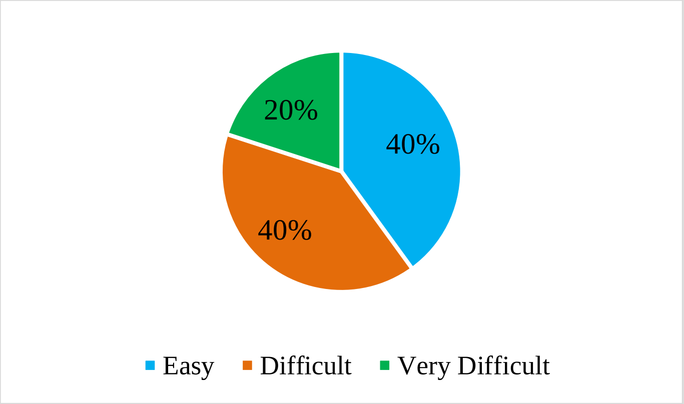

Figure 4. Item difficulty level conceptual understanding instrument.

The distribution of the difficulty level of the concept comprehension items in Figure 4 shows a balanced proportion with a distribution of 40% easy items, 40% difficult items, and 20% very difficult items. This distribution illustrates that the instrument has covered a proportional range of difficulty levels. The next test is the instrument's discrimination test. The results are presented in Table 15.

Table 15. Analysis of item discrimination level.

No	Discrimination	Criteria
1	0.34	Moderate
2	0.44	Good
3	0.62	Good
4	0.44	Good
5	0.58	Good
6	0.32	Moderate
7	0.27	Enough
8	0.26	Enough
9	0.29	Enough
10	0.35	Moderate

Based on the results of the analysis in Table 15, the concept comprehension test items had a discrimination index ranging from 0.26 to 0.62. Items 2, 3, 4, and 5 were classified as good, while items 1, 6, and 10 were classified as moderate. Meanwhile, items 7, 8, and 9 are classified as fair, so revisions are recommended. Discrimination index indicates the extent to which an item can distinguish between participants with high and low abilities; a value of ≥0.40 is generally considered good, while a value of <0.30 tends to require improvement [26]. The research [32] states that good item discrimination indicates how well a question distinguishes between high- and low-ability students. The discrimination test results can distinguish between students who have understood the competency and those who have not. The following is the percentage of item discrimination presented in Figure 5.

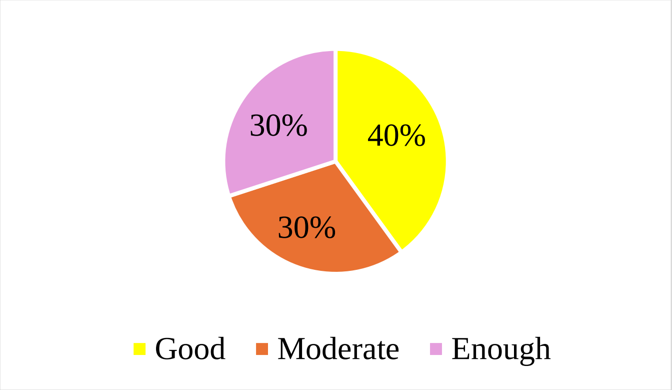

Figure 5. Analysis of item discrimination level.

Figure 5 shows an evenly distributed distribution of power levels. This indicates that the concept comprehension questions are of good quality in measuring the concept comprehension construct in students.

3.4. *Analysis of Psychomotor Assessment Instruments*

The psychomotor assessment instrument developed was an assessment of communication skills through classroom presentations. This assessment was conducted using direct observation to see how skilled the students were in giving presentations. Before the instrument was tested, it had to undergo content validity testing, which in this case involved two physics teachers at a public high school in Yogyakarta. The content validity results can be seen in Table 7. The next step was to conduct a limited trial with 33 students in class XI F4 at SMAN 1 Depok. The data obtained was then analyzed using the Rasch model item testing method to see the characteristics of the items.

3.4.1. Item validity

The first test conducted was the item validity test. The results of this test are presented in Table 16 below:

Table 16. Validity presentation assessment instrument validity test.

Indicator	Infit MNSQ	Description	Decision
Indicator 1	0.71	Less than ideal, but still tolerable.	Suitable for using
Indicator 2	0.92	Fit, Rasch Model	Suitable for using
Indicator 3	1.05	Fit, Rasch Model	Suitable for using
Indicator 4	0.57	Less than ideal, but still tolerable.	Suitable for using
Indicator 5	1.00	Fit ideal	Suitable for using

The results of the analysis presented in Table 16 indicate that four indicators (2, 3, 5) are within the fit range according to the Rasch model. Two indicators (1 and 4) have values below the lower limit (0.71 and 0.57), but this is still tolerable because they do not show excessive deviation. These values are still tolerable because they do not indicate extreme misfit that could interfere with the instrument's construct. [12] explains that items with MNSQ scores outside the ideal range do not necessarily have to be eliminated, but need to be reviewed further based on the context and pattern of respondents' answers. Similar findings were reported by [33] In the vector understanding instrument analysis, several items with MNSQ approaching the lower limit were still deemed acceptable. Thus, the skill assessment instrument indicators were deemed acceptable for use with a review of indicators 1 and 4. The next test is the instrument reliability test to see how reliable/consistent the skill assessment instrument is in measuring the presentation process of students.

3.4.2. Item reliability

Table 17 presents the results of the reliability test of a presentation skill assessment instrument.

Table 17. Reliability of presentation assessment instruments**.**

Mean	1.56
SD	1.1
SD (Adjusted)	0.92
Reliability of estimate	0.74

The reliability value of the presentation skills instrument tested using the Rasch model item method is 0.74. When viewed in the reliability criteria range in Table 1, the reliability value of the skills instrument is classified as reliable because it is in the range of 0.67–0.80. This finding is in line with research [12] which confirms that reliability ≥ 0.70 is considered adequate in exploratory research. Thus, a reliability of 0.74 indicates that the indicators in the presentation skills instrument used in this study are reliable and dependable in measuring students' presentation skills.

After analyzing the three assessment domains, the next step is to measure the effectiveness of using cognitive assessment instruments by comparing the pretest and posttest results in assessing concept understanding. The tests used are the N-Gain statistical test and the effect size. The results obtained are an effect size value of 2.24, which is categorized as a very large effect size. The data shows that there was a very large effect before and after the treatment, which was the implementation of an assessment instrument based on the local wisdom of Kotagede silver crafts.

4. Conclusion

It has been tested and found that the instrument developed as preliminary research is of good quality. This instrument is designed to comprehensively measure affective, cognitive, and psychomotor aspects in physics learning based on the local wisdom of Kotagede silver craftsmanship. The validity and reliability test results show interesting findings. The affective instrument has excellent content validity, but moderate reliability. This means that the indicators used are relevant for measuring attitudes, although the consistency of the assessment needs to be improved. The cognitive instrument shows excellent validity in terms of content, construct, and language, but its reliability is low. This indicates the need for improvements in the variety and number of questions so that the measurements are more consistent. Meanwhile, the psychomotor instruments showed very high validity and reliability. Analysis using the Rasch model showed good fit between the items and individual abilities.

Acknowledgement

The author would like to thank all lecturers in the Physics Education Department at Yogyakarta State University for their knowledge, teaching, and support, which enabled the author to complete this article successfully.

References

[1] Siregar R V, Lubis P K D, Azkiah F and Putri A 2024 Peran Penting Pendidikan dalam Pembentukan Sumber Daya Manusia Cerdas di Era Digitalisasi Menuju Smart Society 5.0 *IJEDR Indones. J. Educ. Dev. Res.* **2** 1408–18

[2] Amir F, M M and Nur A A S 2024 Pengembangan Instrumen Tes Higher Order Thinking Skills (Hots) Peserta Didik Pada Mata Pelajaran Fisika *J. Fis. dan Pembelajarannya* **6** 86–96

[3] Ekawatiningsih P 2015 Pengembangan Instrumen Penilaian Berbasis Kompetensi Untuk Meningkatkan Kualitas Pembelajaran Produktif di SMK *invotec* **11**

[4] Adams W K and Wieman C E 2015 Analyzing the many skills involved in solving complex physics problems *Am. J. Phys.* **83** 459–67

[5] Setiadi H 2016 Pelaksanaan penilaian pada Kurikulum 2013 *J. Penelit. dan Eval. Pendidik.* **20** 166–78

[6] Mukti H, Suastra I W and Aryana I B P 2022 Integrasi Etnosains dalam pembelajaran IPA *JPGI (Jurnal Penelit. Guru Indones.* **7** 356–62

[7] Murti W W and Sunarti T 2021 Pengembangan Instrumen Tes Literasi Sains Berbasis Kearifan Lokal Di Trenggalek *ORBITA J. Kajian, Inov. dan Apl. Pendidik. Fis.* **7** 33

[8] Sumarni W, Sudarmin, Wiyanto and Supartono 2016 The reconstruction of society indigenous science into scientific knowledge in the production process of palm sugar *J. Turkish Sci. Educ.* **13** 281–92

[9] Angraeni D N, Suherman A and Guntara Y 2020 Aplikasi Rasch Model: Pengembangan Fluids Assessment (Fass) Berdasarkan Taxonomy of Introductory Physics Problems (TIPP) *J. Penelit. Pembelajaran Fis.* **11** 135–43

[10] Cholifah S N and Novita D 2022 Pengembangan E-LKPD Guided Inquiry-Liveworksheet untuk Meningkatkan Literasi Sains pada Submateri Faktor Laju Reaksi *Chem. Educ. Pract.* **5** 23–34

[11] Sugiyono 2019 *Metode Penelitian Kuantitatif, Kualitatif, dan R&D* (Bandung: Alfabeta)

[12] Bond T G and M.Fox C 2015 *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (New York: Routledge Taylor & Francis Group)

[13] Gliem J and Gliem R 2003 Calculating, Interpreting, And Reporting Cronbach’s Alpha Reliability Coefficient For Likert-Type Scales *2003 Midwest Res. to Pract. Conf. Adult, Contin. Community Educ.*

[14] Didik Setyawarno 2017 Upaya Peningkatan Kualitas Butir Soal Dengan Analisis Aplikasi Quest *Anal. Biochem.* **11** 1–5

[15] Robert L. Ebel D A F 1986 *Essentials of Educational Measurement* (Amerika Serikat: Prentice-Hall)

[16] Cohen J 1988 *Statistical Power Analysis for the Behavioral Science. In Lawrence Erlabaum Associates (2nd Edition).* (New York: Lawrence Erlabaum Associates)

[17] Rampean B A O and Rohaeti E 2025 The development of an integrated instrument to measure higher order thinking skills and scientific attitudes *J. Turkish Sci. Educ.* **22** 48–62

[18] Gendenjamts S 2023 Measuring Higher-Order Thinking Skills in Science Among Primary School Students Using Item Response Theory *Eur. J. Educ. Stud.* **10** 19–28

[19] Kurniawan K, Sinaga P and Saepuzaman D 2025 Aiken's V analysis and Rasch modeling to determine the quality of physics creative thinking test instruments *J. Ris. dan Kaji. Pendidik. Fis.* **12** 32–40

[20] Fadilah N and . M 2019 Design and Content Validity Analysis of Physics Test based on Local Wisdom for High School Students *Int. J. Educ. Res. Rev.* **4** 602–9

[21] Sari D K and Supahar 2018 The Content Validity of Assessment Instruments to Measure Analogical-Transfer Ability *Int. Jorunal Sci. Basic Appl. Res.* **39** 165–72

[22] Mardikawati R A and Mundilarto M 2020 Development of Physics Communication Skill Instruments Based on Local Wisdom for Senior High School Students *JPI (Jurnal Pendidik. Indones.* **9** 236

[23] Diniya D, Muslim M, Rusdiana D, Permana N D, Andriani R, Sufarman A, Putri M D, Hermita N and Nuraeni F 2024 *Analysis of the Aiken Index in the Development of Scientific Argumentation Written Test on Fluid Mechanics Course* (Atlantis Press SARL)

[24] Riduwan 2012 *Skala Pengukuran Variabel-Variabel Penelitian* (Jakarta: Alfabeta)

[25] Arikunto S 2021 *Dasar-Dasar Evaluasi Pendidikan Edisi 3* (Jakarta: PT BUMI AKSARA)

[26] Ebel and Frisbie 2009 *Essentials Of Educational Measurement 5Th Ed.* (Prentice-Hall Of India Pvt. Limited)

[27] Maric D, Fore G A, Nyarko S C and Varma-Nelson P 2023 Measurement in STEM education research: a systematic literature review of trends in the psychometric evidence of scales *Int. J. STEM Educ.* **10**

[28] Singh T 2024 Validity in student assessment: Implications for competency-based curriculum *Indian J. Heal. Sci. Biomed. Res. KLEU* **17** 1–4

[29] Menon V, Grover S, Gupta S, Indu P V., Chacko D and Vidhukumar K 2025 A primer on reliability testing of a rating scale *Indian J. Psychiatry* **67** 725–9

[30] Hardianti H, Liliawati W and Tayubi Y R 2023 Karakteristik tes kemampuan berpikir kritis siswa SMA pada materi momentum dan impuls: Perbandingan classical theory test (CTT) dan model Rasch *WaPFi (Wahana Pendidik. Fis.* **8** 21–8

[31] Arikunto S 2021 *Dasar-Dasar Evaluasi Pendidikan Edisi 3* (Jakarta: PT BUMI AKSARA)

[32] Adawiyah R and Wisudawati A W 2017 Pengembangan Instrumen Tes Berbasis Literasi Sains : Menilai Pemahaman Fenomena Ilmiah Mengenai Energi *Indones. J. Curric.* **5** 112–21

[33] Susac A, Planinic M, Klemencic D and Milin Sipus Z 2018 Using the Rasch model to analyze the test of understanding of vectors *Phys. Rev. Phys. Educ. Res.* **14** 1–6