Phi: Jurnal Pendidikan Fisika dan Terapan
Vol.12 No.1 (2026) 54-65
ISSN: 2460-4348, e-ISSN: 2549-7162
https://jurnal.ar-raniry.ac.id/index.php/jurnalphi/index

Article Type: *original research*

Evaluating Assessment Instruments for Measuring Students' Higher-Order Thinking Skills in Heat and Temperature Concepts

Jessica Novaisya Sinurat[1], Safriana[1*], Desy Sary Ayunda[1], Fajrul Wahdi Ginting[1], Riza Andriani[1]
[1]Department of Physics Education, Faculty of Teacher Training and Education, Malikussaleh University, Aceh, Indonesia

Correspondence E-mail: safriana@unimal.ac.id

ARTICLE INFO

Article History:
Received: 08 October 2025
Revised: 02 December 2025
Accepted: 01 January 2026
Published: 18 January 2026

Keywords:
HOTS instruments; Temperature and heat; Borg & Gall

ABSTRACT

This study aimed to develop a higher-order thinking skills (HOTS)–based test instrument on temperature and heat topics and to examine its initial quality and feasibility as an assessment tool. The research employed a research and development (R&D) approach using the Borg and Gall model, which included stages of potential problem identification, information collection, product design, expert validation, product revision, and limited field testing. Expert validation using Aiken's V yielded coefficients ranging from 0.82 to 0.94, indicating strong content validity. The limited field trial involved 32 ninth-grade students, and data analysis was conducted using Microsoft Excel and IBM SPSS Statistics 25. Empirical validity testing showed that 23 out of 40 items (57.5%) met the validity criterion, while the remaining items required revision or elimination. Reliability analysis using the KR-20 formula produced a coefficient of 0.776, indicating high internal consistency. Analysis of item difficulty revealed that 10 items (25%) were classified as difficult and 30 items (75%) as moderately difficult. Furthermore, the discriminating power analysis indicated that only 10 items (25%) demonstrated good discrimination, while 15 items (37.5%) fell into the sufficient category, suggesting that several items require further refinement to better differentiate students' ability levels. Overall, the findings indicate that the developed HOTS-based test instrument demonstrates preliminary quality and conditional feasibility for measuring students' higher-order thinking skills in junior high school science learning on temperature and heat topics. However, substantial revisions—particularly to improve item discrimination and increase the proportion of valid items—are necessary before the instrument can be recommended for broader implementation.

1. INTRODUCTION

Higher-order thinking skills (HOTS) are widely recognized as essential competencies in science education, as they enable students to analyze information, evaluate evidence, and create solutions to contextual problems. In the context of junior high school science learning, HOTS play a crucial role in fostering conceptual understanding and problem-solving abilities, particularly in abstract topics such as temperature and heat(Hamdan et al., 2024; Kumala et al., 2024). However, empirical studies consistently report that students' HOTS levels remain relatively low, especially in science subjects. One contributing factor is the dominance of assessment practices that emphasize memorization and basic comprehension rather than higher-order cognitive processes (Ambarwati & Dasari, 2022; Ichsan et al., 2019). Although HOTS have been formally integrated into the curriculum, many teachers experience difficulties in translating HOTS indicators into valid and reliable assessment instruments. These difficulties include limited understanding of HOTS operationalization, lack of validated item examples, time constraints, and challenges in aligning assessment tasks with students' cognitive levels (Aini et al., 2023; Hartono et al., 2022; Supriyadi et al., 2022).

This condition was also identified through preliminary interviews with science teachers at SMP Negeri 5 Lhokseumawe. The findings revealed that while teachers are familiar with the concept of HOTS, they rarely employ HOTS-based test instruments in classroom assessment. Most evaluation tools used in science learning still focus on lower-order thinking skills due to the absence of locally available, validated HOTS question banks and the limited opportunities for teachers to collaboratively develop and test assessment instruments. Consequently, students have minimal exposure to HOTS-oriented questions, which may hinder the development of their higher-order cognitive abilities.

The challenge is particularly apparent in the assessment of temperature and heat materials. These topics require students to interpret data, analyze relationships between variables, and apply concepts to everyday phenomena. Nevertheless, previous studies indicate that assessment instruments for temperature and heat often fail to incorporate contextual problems, graphical representations, or open-ended reasoning tasks that reflect HOTS demands (Handayani et al., 2019; Sarah et al., 2022). As a result, existing instruments may not adequately capture students' higher-order thinking performance in this domain. Although several studies have discussed HOTS conceptually or examined students' HOTS profiles, research focusing on the systematic development and empirical evaluation of HOTS-based test instruments particularly for temperature and heat topics at the junior high school level remains limited. Most available instruments have not been rigorously analyzed in terms of validity, reliability, item difficulty, and discriminating power within authentic classroom contexts. herefore, this study aims to develop a HOTS-based test instrument on temperature and heat materials and to examine its validity, reliability, difficulty level, and discriminating power through a limited field trial. The findings are expected to provide empirical evidence to support the improvement of assessment practices and to offer a reference for teachers in implementing HOTS-oriented evaluation in science learning.

2. METHODS

This study employed a Research and Development (R&D) approach adapted from the Borg and Gall development model, as described by (Sugiyono, 2016). The original Borg and Gall model consists of ten sequential stages, namely: (1) research and information collecting, (2) planning, (3) develop preliminary form of product, (4) preliminary field testing, (5) main product revision, (6) main field testing, (7) operational product revision, (8) operational field testing, (9) final product revision, and (10) dissemination and implementation. In this study, the Borg and Gall model was modified and simplified to suit the research objectives, scope, and time constraints. From the original ten stages, only six stages were implemented, while the remaining four stages were intentionally omitted. The implemented stages were as follows (Gustina et al., 2024):

1) Research and information collecting (identification of potential problems and needs related to HOTS-based assessment);
2) Planning (determination of HOTS indicators, test specifications, and blueprint design);
3) Develop preliminary form of product (construction of HOTS-based test items on temperature and heat materials);
4) Expert validation (content validation by subject-matter and assessment experts);
5) Product revision (revision of test items based on expert feedback);
6) Limited field testing (empirical testing of the instrument with a small group of students)

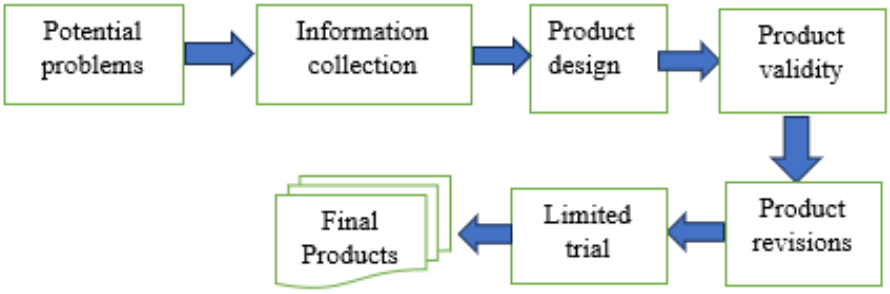

Figure 1. Borg & Gall Steps

Furthermore, a development flowchart illustrating the modified R&D procedure is provided in Figure 1 to clarify the sequence of activities conducted in this study. It should be noted that, due to the restriction of development stages to a limited field trial, the findings of this study represent conditional and preliminary evidence of the instrument's quality. Further studies are recommended to continue the remaining Borg and Gall stages to strengthen the validity, reliability, and practical applicability of the instrument across broader educational contexts.he test subjects of the test instruments developed in this study are 32 students in grades IX-1 of SMP Negeri 5 Lhokseumawe who have studied temperature and heat materials.

2.2 Data Analysis Techniques

2.2.1 Content validity

The validity of the content is determined from the agreement of experts by looking at the content/material, construction, and language (Hidayah & Muhtarom, 2023). The Aiken formula (Aiken, 1985; An Nabil et al., 2022) is used to calculate and determine the validity coefficient (V) of rating-scale data, formulated as follows:

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

Information:

V = Aiken's V validity index
r = score given by the validator
l_0 = lowest score on the rating scale
c = highest score on the rating scale
s = score given by the validator minus the lowest score ($s=r-l_0$)
n = number of validators

2.2.2 Empirical Validity Test

The validity of the calculation was obtained using the product moment correlation formula of the gross number.

$$r_{XY} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{\{N\sum X^2 - (\sum X)^2\}\{N\sum Y^2 - (\sum Y)^2\}}} \quad (2)$$

Information:

r_{XY} = the correlation coefficient of the variables X and Y
N = Number of test subjects
$\sum x$ = Total item score (x)
$\sum y$ = Total item score (y)
$\sum x^2$ = Number of rank score items (x)
$\sum y^2$ = Number of item score rank (y)
$\sum xy$ = Sum of the multiplication of the total score (x) and the total score (y)

Table 1. Validity Criteria

Correlation Coefficients	Criterion
$0{,}81 < r_{xy} \leq 1{,}00$	Very high
$0{,}61 < r_{xy} \leq 0{,}80$	Tall
$0{,}41 < r_{xy} \leq 0{,}60$	Keep
$0{,}21 < r_{xy} \leq 0{,}40$	Low
$0{,}00 < r_{xy} \leq 0{,}20$	Very low

2.2.3 Reliability Test

The reliability test aims to ensure that the instrument produces consistent data if repeated under the same conditions or where the instrument is called reliable if used multiple times to measure the same object, it will produce the same data. Calculating the reliability of the instrument used the Richardson Kuder formula of 20 (KR-20) (Arikunto, 2018).

$$r_{11} = \frac{n}{n-1}\left[\frac{s^2 - \sum pq}{s^2}\right] \quad (3)$$

Information:

r_{11} = overall test reliability coefficient
n = number of test items
p = proportion of respondents who answered an item correctly
q = proportion of respondents who answered an item incorrectly (q=1−p)
Σpq = sum of the products of p and q for all items
s^2 = variance of the total test scores

Table 2. Test Item Reliability Criteria

Reliability coefficient	Reliability criteria
$0.80 < r_{11} \leq 1.00$	Very high

$0.60 < r_{11} \leq 0.80$	Tall
$0{,}40 < r_{11} \leq 0.60$	Enough
$0{,}20 < r_{11} \leq 0.40$	Low
$0{,}00 < r_{11} \leq 0.20$	Very low

2.2.4 Difficulty Level Test

Test Difficulty Level is a measure that shows how difficult or easy a question or test is for respondents.

$$P = \frac{B}{JS} \tag{4}$$

Information:
P = Difficulty Index
B = Number of test participants answered correct
JS = Number of all test takers

Table 3. Test item difficulty criteria

Correlation coefficient	Difficulty level indicator
$0{,}71 < P \leq 1{,}00$	Easy
$0{,}31 < P \leq 0{,}70$	Keep
$0{,}00 < P \leq 0{,}30$	Difficult

2.2.5 Differentiating Power Test

The differentiating power test is the ability of a question to distinguish between high-ability students and low-ability students (Joko Widiyanto, 2018).

$$D = \frac{BA}{JA} - \frac{BB}{JB} \tag{5}$$

Information:
D = item differentiating power index
BA = number of students in the upper group who answered the item correctly
BB = number of students in the lower group who answered the item correctly
JA = number of students in the upper group
JB = number of students in the lower group

Table 4. Differentiating Power index criteria

Differentiating power coefficient	Differentiating power criteria
$0{,}71 \leq DP \leq 1{,}00$	Excellent
$0{,}41 \leq DP \leq 0{,}70$	Good
$0{,}21 \leq DP \leq 0{,}40$	Enough
$0{,}00 \leq DP \leq 0{,}20$	Bad
$0 < DP$	Very bad

3. RESULT AND DISCUSSION

3.1 Results

3.1.1 Item Validity Analysis

An item validity analysis was conducted to determine the extent to which each test item measured the intended construct of higher-order thinking skills (HOTS). The analysis employed the Product Moment correlation, which is commonly used to examine the relationship between individual item scores and total test scores in educational measurement.

Prior to the validity analysis, the item score data were examined descriptively to ensure sufficient score variation, which is a basic requirement for correlation-based item analysis. The validity testing was conducted using a sample of 32 students (N = 32). Based on this sample size, the degree of freedom was df = N − 2 = 30, and the critical value of the correlation coefficient (r-table) at a significance level of α = 0.05 was 0.361. An item was considered valid if the obtained correlation coefficient (r-count) was greater than or equal to the r-table value.

Table 1. Results of the Validity of HOTS-Based Test Questions

Information	No. Question Item	Sum
Valid	8, 9, 11, 12, 13, 15, 16, 18, 19, 20, 21, 24, 25, 26, 28, 29, 31, 32, 34, 35, 37, 38, 39	23
Invalid	1, 2, 3, 4, 5, 6, 7, 10, 14, 17, 22, 23, 27, 30, 33, 36, 40	17

The results of the empirical validity analysis indicate that 23 out of 40 test items were declared valid, while 17 items did not meet the validity criterion and were categorized as invalid. These findings suggest that more than half of the developed items demonstrated adequate correlation with the overall test score, indicating acceptable construct representation within the context of this study.

Furthermore, the valid items were distributed across the three HOTS cognitive levels, namely C4 (analysis), C5 (evaluation), and C6 (creation). Specifically, 10 items were categorized under C4, 10 items under C5, and 3 items under C6. This distribution shows that the instrument primarily assesses analytical and evaluative thinking skills, while items measuring creative thinking remain limited. Therefore, although the instrument demonstrates preliminary construct validity, revisions are required to improve the balance and representation of HOTS indicators, particularly at the C6 level.

Table 2. Indicators of Valid Question Items

Indicator HOTS	Question No.	Sum
C4 (analyze)	11, 12, 13, 16, 19, 21, 24, 29, 31, 37	10
C5 (evaluate)	8, 9, 15, 20, 25, 26, 32, 34, 35, 38	10
C6 (creation)	18, 28, 39	3

3.1.1 Reliability

A reliability analysis was conducted to examine the internal consistency of the HOTS-based test instrument. The analysis employed the Kuder–Richardson Formula 20 (KR-20), which is appropriate for instruments consisting of dichotomously scored items. The use of KR-20 is justified by the scoring format of the test and its frequent application in educational assessment research.

The reliability analysis was performed using IBM SPSS Statistics 25. The results show that the instrument obtained a KR-20 reliability coefficient of 0.776, which exceeds the commonly accepted minimum threshold of 0.70, indicating high internal consistency. The reliability classification in this study follows standard measurement criteria, where coefficients between 0.70 and 0.79 are considered high and acceptable for research purposes.

Table 3. Reliability Test Results

Reliability (KR-20)	Criterion
0,776	High

Although the obtained reliability coefficient indicates that the instrument produces consistent measurements under the conditions of this study, the result should be interpreted cautiously. The reliability estimate is limited to the sample size (N = 32) and testing conditions used in this research. Therefore, further reliability testing involving larger and more diverse samples is recommended to strengthen the generalizability and stability of the instrument.

3.1.2 Difficulty Level

The results of the difficulty level in this study can be seen in Table 4.

Table 4. Difficulty Level Results

Criterion	No. Question Item	Sum
Difficult	8, 9, 15, 16, 20, 22, 24, 32, 34, 35	10
Keep	1, 2, 3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 17, 18, 19, 21, 23, 25, 26,27, 28, 29, 30, 31, 33, 36, 37, 38, 39, 40	30
Easy	-	-

The results of the difficulty level test showed that out of 40 questions, there were 30 questions with a medium difficulty level and 10 questions with a difficult difficulty level, while none of the questions were included in the easy category. This proportion shows that the questions developed are in accordance with the characteristics of a good instrument, because most of the questions are at a moderate difficulty level so that they can be answered by students with average ability. The existence of questions with difficult categories is also important to measure the ability of students with a higher level of understanding.

3.1.3 Differentiating Power Test

The differentiating power test is carried out to determine the ability of a question to distinguish between students who master the material and students who do not master the material. Can be seen in the table.

Table 5. Differentiating Power Test Results

Category	No. Question Item	Sum
Excellent	-	-
Good	9, 16, 20, 24, 26,29, 32, 34, 38, 39	10

Enough	5, 6, 8, 11, 12, 13, 15, 18, 19, 21, 25, 28, 31, 35, 37	15
Poor	2, 3, 7, 10, 17, 27, 30, 33, 36, 40	10
Very poor	1, 4, 14, 22, 23	5

Based on the results of the discriminating power analysis, only 10 out of 40 items (25%) were classified in the good category, while 15 items (37.5%) were categorized as fair, 10 items (25%) as poor, and 5 items (12.5%) as very poor. No items reached the very good category. These findings indicate that the majority of the test items (62.5%) demonstrate insufficient discrimination power and therefore have limited ability to differentiate between students with high and low levels of HOTS.

Items classified in the good category may be retained for use without major modification. Items in the fair category, although still usable, require substantial revision to improve their ability to distinguish students' cognitive performance. In contrast, items categorized as poor and very poor are not suitable for use in their current form and should either be extensively revised or replaced, as they fail to function effectively as discriminative assessment tools.

These results suggest that, despite the presence of a limited number of well-performing items, the overall quality of the instrument in terms of discriminating power does not yet meet the criteria for a fully effective assessment instrument. Therefore, the instrument should not be considered ready for broad implementation. Instead, the findings indicate that the instrument demonstrates preliminary quality and conditional feasibility, contingent upon substantial revision of a considerable proportion of test items, particularly those with low discrimination indices.

3.1.4 Teacher's response results

After conducting a limited trial stage, the researcher also distributed a questionnaire of teachers' responses to the HOTS-based test instrument with 10 statements. The questionnaire was distributed to seven science teachers of SMP Negeri 5 Lhokseumawe. This aims to see teachers' responses to the instrument regarding the use of HOTS-based test instruments. Based on the results of the teacher's response, the HOTS-based test instrument on temperature and heat materials is considered very useful for use in schools. Teachers stated that this instrument helps them assess students' high-level thinking skills more precisely, not just limited to the memorization aspect. This instrument can also be used as an example or reference in the preparation of HOTS questions in other materials, so as to support the improvement of the quality of assessment in schools (Koza et al., 2024).

Teachers' responses indicate that the developed HOTS-based test instrument aligns with curriculum demands and is perceived as practical for use in the assessment of temperature and heat materials. The average teacher response score reached 89%, which falls within the very feasible category based on the applied response criteria. This finding suggests that, from a practitioner's perspective, the instrument is considered relevant and applicable in classroom assessment contexts, particularly because it was designed according to HOTS indicators at the C4 (analysis), C5 (evaluation), and C6 (creation) levels. However, further examination of the item composition reveals a notable imbalance in the representation of HOTS indicators. Of the 40 items developed, only three items (7.5%) assess the C6 (creation) level, while the majority of items focus on C4 and C5. This uneven distribution raises concerns regarding the comprehensiveness of the HOTS construct measured by the instrument. Although C4 and C5 are essential components of higher-order thinking, the limited number of C6 items suggests that the instrument does not yet adequately capture students' creative thinking abilities, which constitute a critical dimension of HOTS. Moreover, the teacher response data should be interpreted cautiously. While the high response score reflects positive perceptions of practicality and curricular relevance, it remains subjective and was not triangulated with direct evidence of student performance outcomes. Therefore, teacher responses alone cannot be used as definitive evidence of the instrument's effectiveness in

measuring students' HOTS abilities. This limitation highlights the need for further empirical validation that integrates teacher perceptions with student performance data (Agusta et al., 2019).

3.2 Discussion

The study adapted the simplified Borg & Gall model into six stages due to time constraints. The first stage**,** problem identification, was carried out through interviews with science teachers at SMP Negeri 5 Lhokseumawe. It was found that the questions used were still at a low cognitive level (C1–C3), so HOTS-based instruments (C4–C6) were needed on temperature and heat materials. The second stage**,** information collection, includes the study of curriculum, materials, and previous research as the basis for product preparation. The third stage**,** product design, produces draft HOTS-based questions, grids, and assessment rubrics with contextual questions. In the fourth stage**,** product validation, three experts assess the content, construction, and language aspects. The results showed that the instrument was very valid with an Aiken V value between 0.82–0.94. The fifth stage**,** product revision**,** is carried out based on validators' suggestions, including clarifying stimulus, improving layout, language, and improving visual quality. The sixth stage, a limited trial, was carried out on 32 students (Ananda et al., 2022). The results: 23 questions were valid, high reliability (α = 0.776), moderate dominant difficulty, and good discriminating power. The results of the limited trial produced 23 valid questions, where there were 10 questions C4 (analysis), 10 questions C5 (evaluating) and 3 questions C6 (creation). In addition, the response questionnaire from the teacher showed an average assessment of 89% (very decent). Teachers assess the instruments as appropriate, challenging, and effective for measuring students' high-level thinking skills (Miladanta et al., 2024; Rahmaniasan et al., 2022).

The final product in the form of a HOTS-based test instrument on temperature and heat material produced 23 valid questions, consisting of 10 questions in the C4 category (analysis), 10 questions in the C5 category (evaluation), and 3 questions in the C6 category (creation). In the C4 category, valid questions are at numbers 11, 12, 13, 16, 19, 21, 24, 29, 31, and 37. The results of the analysis show that all of these questions have empirical validity that meets the criteria because the calculated r value is greater than the r table, so that it is able to measure students' analytical ability appropriately. The instrument also has high reliability, with a KR-20 coefficient of more than 0.7 so that it is consistent in measuring. In terms of difficulty, most of the questions are in the medium category, although there are questions that fall into the difficult category, such as numbers 16 and 19 that require students to organize the data of the experiment results. Meanwhile, the differentiating power of the questions is in the category of good and sufficient, which means that the questions are able to distinguish between high-ability and low-ability students. These ten questions require students to be able to break down concepts into small parts, organize data, and find the meaning of a phenomenon. Questions 11–13 emphasize the differences in the concepts of temperature, heat, and questions 21, 24, 29, 31, and 37 practice the ability to provide explanations or reasons, such as finding the purpose of the experiment, the cause of the deviation of the results, or the core meaning of the graph presented. This is in line with the analytical skills that are at the core of C4's cognitive processes.

In the C5 category, valid questions are found in numbers 8, 9, 15, 20, 25, 26, 32, 34, 35, and 38. Empirically valid, all of these questions have been proven to be valid and have a significant relationship with the total score, so they can be used to measure students' evaluative abilities. The instrument's reliability value is also high (KR-20 > 0.7), indicating that the instrument is consistent in measuring(Maryani et al., 2021; Rintayati et al., 2020). From the results of the difficulty level analysis, most of the questions are moderate, while questions number 8 and 9 are included in the easy category because they emphasize the examination of heat calculation results and simple experimental procedures. In terms of differentiating power, most of the questions are in the good and sufficient categories, so it is effective to distinguish students with different levels of evaluative ability. These ten questions focus on the ability to evaluate, both through the process of checking and criticizing. Questions 15 and 20 require students to compare the effectiveness of two heating methods or technologies;

Meanwhile, questions number 25, 26, 32, 34, 35, and 38 encourage students to provide scientific criticism of data, graphs, and the use of measuring instruments. Thus, the questions in the evaluation category train students' critical thinking skills in providing assessments based on internal and external criteria.

Meanwhile, in the C6 category, valid questions are at numbers 18, 28, and 39. The results of empirical validity show that these three questions are valid, so they can be used to measure students' creative abilities. In terms of reliability, this question item is consistent with the overall instrument, with a KR-20 value of more than 0.7. Based on the difficulty level analysis, C6 category questions tend to be in the difficult category because they require original and open-ended answers. In terms of differentiation, the questions are in the good category, so they can distinguish students with high and low levels of creativity. These three questions require students' creative skills, namely generating hypotheses, planning, and creating new designs(Hasanah et al., 2020). For example, question number 18 directs students to generalize a phenomenon and formulate a hypothesis; Question number 28 emphasizes the ability to plan simple experiments that prove the relationship between heat and change of form; Meanwhile, question number 39 asks students to create a simple technology design based on the principle of heat transfer. Although there are fewer of them, the C6 category questions are very important because they train students' creativity and innovation which includes the ability to generate hypotheses, plan, and create new products that are relevant to real life.

4. CONCLUSION

The results of this study indicate that the assessment instrument evaluated demonstrates an acceptable initial quality for measuring students' higher-order thinking skills in heat and temperature concepts. Analyses of item validity, reliability, difficulty level, and discriminating power show that the instrument meets basic measurement criteria within the context of the limited trial conducted. Nevertheless, these findings are constrained by the scope of the study and do not yet represent comprehensive empirical evidence of measurement effectiveness. Although teachers' responses yielded a high feasibility score (89%), this result primarily reflects perceived practicality and alignment with curricular demands rather than confirmed measurement accuracy. Accordingly, the instrument should be considered feasible only under certain conditions and requires further refinement. Improvements are particularly needed to enhance construct representation and to strengthen the quality of several test items before the instrument can be recommended for wider application in assessing higher-order thinking skills related to heat and temperature topics.

REFERENCES

Agusta, S., Sitompul, S. S., & Arsyid, S. B. (2019). Pengembangan tes Higher Order Thinking Skill (HOTS) pada materi suhu dan kalor untuk SMA. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa*, *8*(10), 1–13.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Aini, R., Karma, I. N., & Affandi, L. H. (2023). Kesulitan guru dalam menyusun soal evaluasi berbasis Higher Order Thinking Skills dalam pembelajaran Kurikulum 2013. *Jurnal Educatio*, *9*(4), 2062–2069. https://doi.org/10.31949/educatio.v9i4.6035

Ambarwati, D., & Dasari, D. (2022). HOTS-oriented learning model and mathematical reasoning ability. *The Journal of Analysis*, *8*(2), 130–141. https://doi.org/10.15575/ja.v8i2.22322

An Nabil, N. R., Wulandari, I., Yamtinah, S., Ariani, S. R. D., & Ulfa, M. (2022). Analisis indeks Aiken untuk mengetahui validitas isi instrumen asesmen kompetensi minimum berbasis konteks

sains kimia. *Paedagogia: Jurnal Penelitian Pendidikan*, *25*(2), 184–192. https://doi.org/10.20961/paedagogia.v25i2.64566

Ananda, D., Kartono, K., & Ghasya, D. A. V. (2022). Pengembangan instrumen soal High Order Thinking Skills pada pembelajaran tematik kelas V sekolah dasar. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK)*, *11*(2), 1–10. https://doi.org/10.26418/jppk.v11i2.52406

Arikunto, S. 2018. *Dasar-dasar Evaluasi Pendidikan* (R. Damayanti, Ed.). PT Bumi Aksara.

Hartono, A., Hasairin, A., & Setya Diningrat, D. (2022). Development of HOTS-based questions on biology learning. *Jurnal Biolokus: Jurnal Penelitian Pendidikan Biologi dan Biologi*, *5*(1), 54–65. https://doi.org/10.30821/BIOLOKUS.V5I1.1351

Gustina, Z., Husnayayin, A., Eka, D., & Dewi, C. (2024). Karakteristik dan langkah-langkah metode penelitian Research and Development (Borg & Gall) dalam pendidikan. *Jurnal Ilmiah Pendidikan Dasar*, *9*, 490–501.

Hamdan, N., Heong, Y. M., Masran, S. H., Kiong, T. T., Sutadji, E., & Fuada, S. (2024). Exploring Marzano Higher-Order Thinking Skills: Demographic disparities among technical students. *Journal of Technical Education and Training*, *16*(2), 103–118. https://doi.org/10.30880/JTET.2024.16.02.009

Handayani, F., Hartono, H., & Lestari, W. (2019). Need analysis in the development of HOTS-oriented study project assesment instrument in android-based science learning. *Journal of Research and Educational Research Evaluation*, *8*(1), 57–64. https://doi.org/10.15294/JERE.V8I1.31799

Hasanah, T. A. N., Handayanto, S. K., Zulaikah, S., & Yuenyong, C. (2020). How are student's cognitive patterns viewed from higher-order thinking skills in kinematics? *Journal of Physics: Conference Series*, *1567*(3), Artikel 032077. https://doi.org/10.1088/1742-6596/1567/3/032077

Hidayah, N., & Muhtarom. (2023). Validity and reliability test of teaching materials using Aiken's V formula and SPSS 22. *Schola Journal: Jurnal Ilmiah Pendidikan Dasar*, *1*(2), 75–82. https://doi.org/10.26877/schola.v1i2.342

Ichsan, I. Z., Sigit, D. V., Miarsyah, M., Ali, A., Negeri, U. I., Makassar, A., & Prayitno, T. A. (2019). HOTS-AEP: Higher order thinking skills from elementary to master students in environmental learning. *European Journal of Educational Research*, *8*(4), 935–942. https://doi.org/10.12973/eu-jer.8.4.935

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

Kumala, F. N., Safitri, I., Triwahyudianto, T., Yasa, A. D., & Salimi, M. (2024). HOTS-based e-evaluation Quizwhizzer in science learning in elementary schools. *Inovasi Kurikulum*, *21*(3), 1345–1358. https://doi.org/10.17509/JIK.V21I3.71147

Maryani, I., Prasetyo, Z. K., Wilujeng, I., Purwanti, S., & Fitrianawati, M. (2021). HOTs multiple choice and essay questions: A validated instrument to measure higher-order thinking skills of prospective teachers. *Journal of Turkish Science Education*, *18*(4), 674–690. https://doi.org/10.36681/tused.2021.97

Miladanta, A. N., Nuryantini, A. Y., Farida, I., & Cahyanto, T. (2024). Pengembangan instrumen Higher Order Thinking Skills (HOTS) materi alat optik melalui validitas, reliabilitas, daya pembeda dan tingkat kesukaran. *Jurnal Penelitian Sains dan Pendidikan (JPSP)*, *4*(2), 148–158. https://doi.org/10.23971/jpsp.v4i2.7828

Rahmaniasan, V., Berlian, L., Indah Suryani, D., & Sultan Ageng Tirtayasa, U. (2022). Pengembangan instrumen tes two-tier multiple choice berbasis HOTS tema pemanfaatan gelombang untuk menumbuhkan kemampuan berpikir tingkat tinggi siswa. *Jurnal Pendidikan MIPA*, *12*(3), 929–935. https://doi.org/10.37630/JPM.V12I3.706

Rintayati, P., Lukitasari, H., & Syawaludin, A. (2020). Development of two-tier multiple choice test to assess Indonesian elementary students' higher-order thinking skills. *International Journal of Instruction*, *14*(1), 555–566. https://doi.org/10.29333/IJI.2021.14133A

Sarah, S., Aswita, D., Ainun, N., Maulidar, M., & Azzarkasyi, M. (2022). The development of HOTS-based assessment instruments on educational statistics. *International Journal of Trends in Mathematics Education Research*, *5*(1), 38–43. https://doi.org/10.33122/IJTMER.V5I1.107

Sugiyono. (2016). *Metode penelitian kuantitatif, kualitatif, dan R&D*. Alfabeta.

Supriyadi, S., Astuti, N., Sukamto, I., Destini, F., Khairani, F., & Izzatika, A. (2022). Implementation of HOTS-oriented problem based learning on science literacy ability. *JPP (Jurnal Pendidikan Progresif)*, *12*(3), 1492–1499. https://doi.org/10.23960/jpp.v12.i3.202236

Widiyanto, J. (2018). *Evaluasi pembelajaran* (A. Musandi, Ed.). UNIPMA Press.