JPPIPA 11(5) (2025)

Jurnal Penelitian Pendidikan IPA

http://jppipa.unram.ac.id/index.php/jppipa/index

Evaluation of Higher-Order Thinking Skills of Middle School Students on Vibration and Wave Topic Using Rasch Measurement

Annur Indra Kusumadani[1*], Harry Afandy[2], Lina Agustina[1], Rina Astuti[1], Muhamad Waluyo[1]

[1]Faculty of Teacher Training and Education, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia.
[2]Department of Natural Science Education, Universitas Sebelas Maret, Surakarta, Indonesia.

Received: January 26, 2025
Revised: March 17, 2025
Accepted: May 25, 2025
Published: May 31, 2025

Corresponding Author:
Annur Indra Kusumadani
aik120@ums.ac.id

DOI: 10.29303/jppipa.v11i5.10900

Abstract: Higher-order thinking skills (HOTS) are crucial in science education, yet students often struggle to master them, particularly in complex topics such as vibrations and waves. This study aimed to evaluate the HOTS performance of 165 eighth-grade students from five schools in Purbolinggo Sub-district, East Lampung, using a two-tier multiple-choice test analyzed through the Rasch Measurement Model. The test consisted of 12 items, each with a content question and a reasoning tier, adapted from Treagust (1988) and validated through expert judgment. Rasch analysis indicated that most students demonstrated low HOTS ability, especially in the creating domain (C6), involving generating novel ideas or solutions. In contrast, analyzing (C4) and evaluating (C5) were relatively easier but still reflected limited mastery. The highest difficulty appeared in questions requiring divergent thinking and complex synthesis. Item fit statistics showed good instrument validity (person reliability = 0.93; item reliability = 0.83). These findings highlight the importance of implementing learning strategies that foster creativity and synthesis, such as problem-based learning or project-based approaches. The study was limited to one geographic area, which may restrict generalizability. Future research should explore broader contexts and incorporate qualitative data to deepen understanding of HOTS development in science.

Keywords: HOTS; Learning evaluation; Rasch model; Two-tier test; Vibration and wave

Introduction

Higher-order thinking skills (HOTS) a crucial role in junior high school science learning, as they enable students to build conceptual understanding and solve complex problems independently. International assessments such as PISA (2023) have shown that Indonesian students continue to struggle with science items that require analysis, evaluation, and problem-solving, indicating persistent gaps in HOTS mastery. Although multiple studies highlight the value of HOTS in science education (Kwangmuang et al., 2021; Ramadan et al., 2023; Susantini et al., 2022), challenges remain in assessing these skills effectively—particularly in specific science domains such as Vibrations and Waves, where students must apply abstract concepts and reasoning beyond memorization.

Research by Khaeruddin et al. (2023) revealed that students encounter significant difficulties when tackling HOTS-oriented items in Vibrations and Waves due to the cognitive demand of interpreting and applying conceptual relationships. This highlights the importance of valid, reliable, and domain-specific assessment tools that go beyond traditional multiple-choice formats.

How to Cite:
Kusumadani, A. I., Affandy, H., Agustina, L., Astuti, R., & Waluyo, M. (2025). Evaluation of Higher-Order Thinking Skills of Middle School Students on Vibration and Wave Topic Using Rasch Measurement. *Jurnal Penelitian Pendidikan IPA*, *11*(5), 74–84. https://doi.org/10.29303/jppipa.v11i5.10900

While conventional multiple-choice questions are often criticized for focusing on factual recall, well-constructed alternatives—such as two-tier tests—can better capture students' reasoning processes (Istiyono et al., 2018; Treagust, 1988). Two-tier items provide insights into both the answer and the justification, enabling a more nuanced analysis of student understanding and misconceptions.

More accurate assessment instruments, such as the Two-Tier Test and Rasch Measurement, are important in evaluating students' HOTS because conventional assessment methods are often unable to identify the level of conceptual understanding and misconceptions of students in depth. According to research conducted by Treagust (1988) many students can choose the correct answer in multiple-choice questions. Still, they cannot explain their scientific reasoning appropriately, making it difficult to measure HOTS accurately. Moreover, research by Istiyono et al. (2019) showed that many HOTS assessments are still subjective and do not have high reliability in measuring students' thinking skills. A study by Kaltakci et al. (2016) suggests that the use of the two-tier test in science evaluation can improve the accuracy of HOTS measurement and help teachers understand students' thinking patterns. Furthermore, Rasch Measurement is used to analyze assessment results by considering the difficulty level of questions and individual abilities so that it can provide a more objective overview of students' HOTS skills (Hidayatullah et al., 2022). Bond et al. (2021) research indicates that Rasch Measurement provides more accurate estimates than classical methods in measuring HOTS.

Traditional assessment instruments have limitations in measuring HOTS because they tend only to test the ability to remember and understand concepts superficially without evaluating the ability to analyze, synthesize, and solve problems in depth. According to research conducted by Istiyono (2016), many exam questions used in schools still focus on low-level thinking skills (LOTS), such as memorizing formulas or identifying concepts, making them less effective in measuring students' HOTS. Furthermore Fensham & Bellocchi (2013) study found that students who are accustomed to LOTS-based questions have difficulty when facing HOTS questions because they are not used to thinking critically and reflectively. The reason for this limitation is that traditional questions, especially multiple-choice ones, only require students to choose the correct answer without explaining the reasoning or applying the concept in a different context (Jansen & Meoller, 2022; Negara et al., 2024). Therefore, there is a need for innovation in the evaluation system by incorporating alternative assessment techniques, such as two-tier tests or Rasch measurement, in order to measure HOTS more accurately without sacrificing efficiency in the assessment process. However, teacher capacity to construct such instruments remains limited, especially in specific topics such as Vibrations and Waves (Lewis & Smith, 1993; Seibert, 2021).

The main objective of this research is to develop and test a Rasch Model-based assessment instrument that is more accurate in measuring HOTS of junior high school students on Vibration and Wave materials. The present study contributes to the development of assessment instruments using the Rasch Model, which allows for more objective quantitative analysis based on the difficulty level of the questions and students' abilities. The Rasch Model can overcome the weaknesses of classical assessment methods, such as inconsistency in the level of difficulty of questions and bias in assessment. According to Bond et al. (2021) this approach can produce a more accurate and reliable measurement scale, thus providing a more valid mapping of students' HOTS. Furthermore, the present study uses a two-tier test, which not only assesses students' answers but also the reasoning behind their choices, thus identifying students' misconceptions and level of understanding in more depth.

The novelty of this research lies in combining the two-tier test format with Rasch analysis to develop an instrument specifically targeted at measuring HOTS in the context of Vibrations and Waves—a topic that is cognitively demanding yet underrepresented in HOTS research. Unlike previous studies that broadly discuss HOTS, this study provides a focused diagnostic tool for one of the most conceptually challenging areas in middle school physics. Therefore, this study aims to fill the gap by developing and validating a two-tier HOTS assessment instrument based on the Rasch Model for the topic of Vibrations and Waves. The research is important to ensure that HOTS evaluation in science is not only accurate and objective but also reflective of students' real conceptual understanding. This study seeks to explore several key questions. First, it investigates the psychometric characteristics of the two-tier test items developed based on the Rasch Measurement Model in measuring students' HOTS on Vibrations and Waves material. Additionally, it examines how students' HOTS are distributed according to the Rasch analysis results. Finally, it evaluates the overall fit of the Rasch Model in assessing HOTS within this specific topic area.

Method

Study Design

This study employed a quantitative approach with an evaluative research method to measure students'

Higher-Order Thinking Skills (HOTS) on the topics of Vibration and Waves. The Rasch Measurement Model was used to analyze the psychometric properties of items in the Two-Tier Test instrument and to objectively map the distribution of students' abilities. Unlike classical test theory, the Rasch Model addresses issues such as item bias and inconsistency in measuring ability. The use of a Two-Tier Test enabled the researchers not only to evaluate final responses but also to explore students' reasoning, providing insight into their conceptual understanding.

Research Location and Subjects

The research was conducted in Purbolinggo Sub-district, East Lampung Regency. Participants were 165 Grade VIII students from 12 classes across five junior high schools. There were 82 male students (49.70%) and 83 female students (50.30%), with most students aged 13 (59.39%) and the remainder aged 14 (40.61%). The school distribution was as follows: School A (33 students), School B (32), School C (38), School D (30), and School E (32). This distribution supports a diverse sample in assessing HOTS on Vibration and Wave concepts.

Research Instruments

The instrument used was a Two-Tier Test. Tier 1 consisted of multiple-choice items, while Tier 2 asked for justifications supporting the Tier 1 responses. The development of this instrument was guided by Treagust (1988), who advocated for Two-Tier Tests as effective tools for probing scientific understanding and uncovering misconceptions. The instrument was designed to assess three HOTS aspects from the revised Bloom's taxonomy: Analysis (C4), Evaluation (C5), and Creation (C6). Each item was mapped to specific sub-aspects, as shown in Table 1.

Table 1. Distribution of the number of question items

HOTS Aspects	HOTS Sub-aspect	Item Number	Total
Analysis	Differentiate	1, 7, 20	9
	Analyze	6, 8, 15, 16	
	Organizing strategy	2, 9	
Evaluation	Arguing	10, 17, 24	8
	Judging	3, 11, 14, 22, 23	
Create	Providing a point of view	4, 5, 12, 18, 21, 25	8
	Produce	19	
	Design	13	

The scoring guidelines are designed to ensure objectivity and consistency in the assessment of HOTS test results. The scoring system uses polytomous scaling with four categories, adapted from the research of (Affandy et al., 2021). It allows differentiation of students' level of understanding, especially in linking multiple choice answers (Tier 1) with the reasons given (Tier 2), thus identifying the level of conceptual understanding and possible misconceptions. Each student's answer is categorized based on a combination of the correctness of the concepts used and the accuracy of the final result obtained. Suppose students give an incorrect answer on the multiple choice and are also incorrect in the reasoning (Category 1). In that case, they get a score of 1, indicating that they do not understand the concept correctly. Suppose students answer correctly on the multiple choice but incorrectly in the reasoning (Category 2). In that case, they get a score of 2, indicating that their final answer is correct but without strong conceptual understanding. Conversely, if students give incorrect answers but correct reasoning (Category 3), they get a score of 3, indicating that they understand the concept correctly but make mistakes in the application. If students answer correctly on multiple choice and correctly in reasoning (Category 4), they get a score of 4, reflecting complete and accurate conceptual understanding. The polytomous scoring approach is superior to dichotomous (true-false) scoring because it provides a more detailed picture of students' level of understanding rather than simply assessing correct or incorrect answers.

Data Collection Procedure

The data collection involved two main stages: preparation and implementation. During the preparation phase, research permits were obtained, and test items were validated by experts using Aiken's V index. A pilot test was administered to 30 students to assess item clarity and discrimination. During implementation, the main test was conducted with a 45-minute time limit to allow students to demonstrate their HOTS fully. Responses were collected in written form.

Data Analysis

Instrument reliability and validity were examined using the Rasch Model. Person reliability indicated consistency across student responses, while item reliability indicated how well items represented the HOTS construct. Item validity was analyzed through infit and outfit mean square (MNSQ) statistics. Acceptable fit was inferred from values near 1.0 (Engelhard & Wang, 2021). Item difficulty was analyzed using logit values, where higher logits indicate more difficult items (Linacre, 2011). The logit scale ranges from negative to positive values with zero as the mean difficulty point (Bond, 2015). Statistical comparisons between HOTS sub-aspects included Correct Answer Rate, Item Difficulty, Standard Error, Infit and Outfit MNSQ, and Point Measure Correlation—all within the Rasch framework.

Result and Discussion

Descriptive Statistics of Student Response

According to the descriptive statistical analysis of students' responses, there were variations in the level of HOTS based on gender and school origin. The students in the low category were more dominant than those in the medium and high categories.

Regarding gender, male students in the high HOTS category were five students (3.03%), nine students (5.45%) in the medium category, and the majority were in the low category, with 68 students (41.21%). Meanwhile, female students had a more even distribution, with 14 students (8.48%) in the high category, 12 students (7.27%) in the medium category, and 57 students (34.55%) in the low category. It indicates that the proportion of female students achieving higher HOTS levels is slightly greater than that of male students. According to school origin, there is a significant difference in the distribution of HOTS.

School A and School B had more students in the high HOTS category than the other schools, with eight students (4.85%) and nine students (5.45%), respectively. Conversely, Schools C, D, and E tended to dominate students in the low category, with School C having 38 students (23.03%), School D with 26 students (15.76%), and School E with 29 students (17.58%) in the low category, and almost no students from these schools in the high category. The results indicate that there are differences in HOTS ability between schools, which may be due to the quality of learning, educational resources, and teaching strategies applied in each school. Furthermore, this result also underlines that there are still challenges in developing HOTS in junior high school students, especially in schools with a predominance of students in the low category.

Table 2. Descriptive statistics of student response

Demo-	Sub	Number of students (N)		
graphics	categories	High (%)	Medium (%)	Lowly (%)
Gender	Male	5 (3.03)	9 (5.45)	68 (41.21)
	Female	14 (8.48)	12 (7.27)	57 (34.55)
School	School A	8 (4.85)	9 (5.45)	16 (9.70)
	School B	9 (5.45)	7 (4.24)	16 (9.70)
	School C	1 (.61)	1 (.61)	38 (23.03)
	School D	0	4 (2.42)	26 (15.76)
	School E	0	1 (.61)	29 (17.58)

Item Fit Analysis with Rasch Model

The results of the item fit analysis with the Rasch Model show that most items have a good level of fit, although some items have infit and outfit mean square (MNSQ) values outside the ideal range. According to the item difficulty values, the level of item difficulty ranges from -0.80 to 1.40 logits, with an average of 0.00 logits and a standard deviation of 0.61 logits. Items with negative values indicate easier items, while items with positive values indicate more difficult items. Item 19 was the most difficult (1.40 logits), while Item 7 was the easiest (-0.80 logits). In terms of fit statistics, most items had infit and outfit MNSQ values within the acceptable range (0.5 to 1.5 logits). Point measure correlation results, most items had positive correlations (≥0.30), indicating that the items correlated with students' abilities consistently. The reliability of the instrument was rated high overall, with a personal reliability of 0.93 and an item reliability of 0.83. The high person reliability value indicates that the instrument can distinguish students' abilities well. In contrast, the good item reliability value indicates that the items have sufficient consistency in measuring students' HOTS.

Analysis of Level of Difficulty of Questions Based on HOTS Aspect

The results of the analysis of the level of difficulty of questions based on HOTS aspects show that questions that measure analyzing and evaluating skills are relatively easier to answer than questions that measure creating skills. Regarding the analysis aspect, the sub-aspects of distinguishing (-0.67 logit) and organizing strategies (-0.67 logit) have the lowest level of difficulty, indicating that students find it easier to identify differences in concepts and develop strategies in solving problems. The analyzing sub-aspect (-0.46 logit) is also classified as easy, although it is slightly more difficult than the other two sub-aspects.

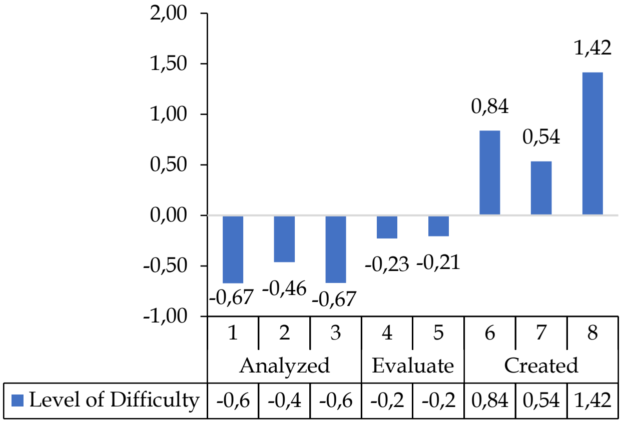

Figure 1. Analysis of the Level of Difficulty of Questions Based on HOTS Aspects

In the evaluating aspect, the sub-aspects of assessing (-0.23 logit) and giving arguments (-0.21 logit) have a difficulty level that is close to the average, indicating that students have a fairly good ability to assess information and provide reasons for a concept.

However, in the aspect of creating, questions that measure the ability to provide a point of view (0.84 logits), design (0.54 logits), and produce (1.42 logits) have a higher level of difficulty than other aspects. The generating sub-aspect (1.42 logits) was the most difficult, indicating that students had difficulty in developing new solutions or generating ideas based on concept understanding. Generally, the results indicate that students find it easier to solve problems that involve identification, analysis, and evaluation but experience greater challenges in developing ideas or creating new solutions.

Table 3. Item fit with Rasch model

Item	Correct answers	Item difficulty	Standar error	Infit MNSQ (logit)	Outfit MNSQ (logit)	Point measure correlation
Item 1	328	-.68	.09	1.40	1.49	.41
Item 2	342	-.78	.09	1.49	1.40	.42
Item 3	301	-.43	.10	.58	.88	.38
Item 4	206	.98	.16	1.32	1.18	.33
Item 5	235	.35	.13	.93	.91	.39
Item 6	285	-.26	.10	1.43	1.47	.36
Item 7	346	-.80	.09	1.00	1.05	.42
Item 8	294	-.34	.10	1.36	1.45	.37
Item 9	323	-.62	.09	.95	1.00	.40
Item 10	294	-.36	.10	.58	.91	.37
Item 11	270	-.11	.11	.76	.85	.34
Item 12	235	.38	.13	.80	.87	.39
Item 13	243	.23	.12	1.39	1.37	.35
Item 14	246	.20	.12	.62	.77	.39
Item 15	327	-.65	.09	1.49	1.42	.41
Item 16	283	-.25	.10	1.36	1.17	.36
Item 17	280	-.22	.10	.76	.65	.36
Item 18	224	.56	.14	1.31	1.29	.37
Item 19	193	1.40	.20	1.26	1.39	.39
Item 20	310	-.51	.09	1.24	1.36	.39
Item 21	209	.87	.16	1.39	1.09	.37
Item 22	283	-.25	.10	.87	.78	.36
Item 23	252	.11	.12	.84	.70	.38
Item 24	261	.00	.11	.85	.82	.33
Item 25	198	1,19	.18	1.43	1.40	.39
M	270.7	.00	.12	1.03	1.02	
SD	44.7	.61	.03	.87	.88	
Person reliability						0.93
Item reliability						0.83

Distribution of Students' Higher Order Thinking Skills

The analysis presents the distribution of students' HOTS based on their ability to score on a polytomous scale. The analysis of students' HOTS distribution indicated that most students obtained low scores (Score 1 and Score 2) in almost all HOTS aspects, especially in the aspect of creating (C6). Regarding the sub-aspect of giving a point of view, most students got a Score of 1, such as in Item 4 (135 students), Item 5 (104 students), and Item 21 (133 students). Similarly, in the sub-aspect of generating (Item 19: 140 students get Score 1), it indicates that students have difficulty in creating or developing new ideas. Meanwhile, in the analyzing aspect (C4), there was a slight increase in students' ability to differentiate concepts and organize strategies, although most students still obtained Score 1 and Score 2. For example, in Item 6 (92 students Score 1) and Item 16 (95 students Score 1), students' analytical skills still need to be strengthened. However, in some items, such as Item 15, as many as 29 students managed to obtain a Score of 4, indicating a small group of students with better analytical understanding.

Furthermore, in the evaluation aspect (C5), most students obtained a score of 2, which indicates that they have sufficient concept understanding but still have difficulty providing appropriate reasons or arguments. Overall, these results indicate that students' higher-order thinking skills are still low, especially in the creating aspect, which requires students to formulate new ideas or provide different points of view. Meanwhile, in the analysis and evaluation aspects, although some students demonstrated better understanding, the majority still had difficulty in providing logical and systematic reasons. Therefore, a learning approach that emphasizes strengthening critical and creative thinking skills, as well as further practice in developing correct concept-based arguments and solutions, is needed.

Discussion

Characteristics of Two-Tier Test Items

The two-tier test instrument has high reliability, as indicated by the person reliability index and item reliability index values, which reflect consistency in measuring students' HOTS. The analysis revealed that the person reliability index value reached 0.93, while the item reliability index was 0.83. The high person reliability value indicates that student's responses to the items are quite consistent and stable. Meanwhile, the item reliability value, which is also quite high, indicates that the instrument has good item quality in measuring HOTS. According to Andrich & Marais (2019), a person's reliability index value above 0.8 indicates that the student's ability measured has a very good level of consistency in the test instrument. Meanwhile, an item reliability index above 0.7 indicates that the items in the instrument have a high level of fit with the Rasch model so that they can provide valid and reliable measurements. The study by Smith (2003) also confirmed that person reliability values above 0.9 indicate that individual responses are very consistent in answering questions, while item reliability above 0.8 indicates that the questions in the instrument have good quality in distinguishing student abilities.

Table 4. Distribution of students' higher order thinking skills

HOTS Aspects	HOTS Sub-aspect	Item Number	Score 1 (N)	Score 2 (N)	Score 3 (N)	Score 4 (N)
Analysis (C4)	Differentiate	1	89	5	55	16
		7	55	43	63	4
		20	78	42	32	13
	Analyze	6	92	39	21	13
		8	98	24	24	19
		15	82	33	21	29
		16	95	42	8	20
	Organizing strategy	2	67	43	31	24
		9	59	63	34	9
Evaluation (C5)	Arguing	10	42	117	6	0
		17	74	72	14	5
		24	71	92	2	0
	Judging	3	41	117	2	5
		11	62	101	2	0
		14	86	77	2	0
		22	51	110	4	0
		23	81	81	3	0
Create (C6)	Providing a point of view	4	135	23	3	4
		5	104	55	3	3
		12	101	60	2	2
		18	117	40	5	3
		21	133	23	6	3
		25	141	16	7	1
	Produce	19	140	23	1	1
	Design	13	117	26	14	8

The validity of the two-tier test instrument in measuring students' HOTS can be determined by analyzing item fit based on Infit Mean Square (MNSQ) and Outfit MNSQ. Items that have Infit and Outfit MNSQ values in the range of 0.5 to 1.5 are considered in accordance with the Rasch model, so they are suitable for measuring HOTS on Vibration and Wave materials. The analysis results show that most of the items have Infit MNSQ and Outfit MNSQ values within the expected range. According to Linacre (2011), Infit and Outfit MNSQ values in the range of 0.5 to 1.5 indicate that the items have a good level of fit with the Rasch model. Items with Infit or Outfit values above 1.5 tend to be less predictable by the model, which can be caused by factors such as student incomprehension or the presence of ambiguous answer alternatives. Studies conducted by Bond (2015) suggest that validity testing with the Rasch model is superior to the classical approach because it considers the relationship between student ability and item difficulty. Furthermore, research by Andrich & Marais (2019) confirmed that items that have Infit and Outfit MNSQ values within the specified range tend to provide a more accurate estimate of students' ability to understand complex scientific concepts. Overall, the validity test results show that most of the items have met the criteria for fit with the Rasch model, so they are suitable for measuring students' HOTS on Vibration and Wave materials.

The items used in the study have good characteristics in measuring students' HOTS, as shown by the item difficulty and item fit analysis. The results of the item difficulty analysis show that the difficulty level of the questions is spread in the range of -0.80 to 1.40 logits, with an average value of 0.00 logits and a standard deviation of 0.61 logits. Most questions had negative logit values, indicating that the questions were easier for students, while some questions with positive logits were more difficult. The fairly varied distribution of question difficulty levels indicates that this instrument is able to accommodate students with different levels of HOTS understanding (Linacre, 2011). Previous research by Zakwandi et al. (2024) showed that two-tier test-based tests with Rasch Model analysis can provide accurate information about students' conceptual understanding. The results of the study are also in line with the findings of Affandy et al. (2021), which show that item difficulty analysis in the Rasch Model can help in developing items that are suitable for students' abilities and increase the validity of HOTS measurement.

Distribution of HOTS to Students

Students' HOTS in the categories of analyzing (C4), evaluating (C5), and creating (C6) indicate a diverse distribution pattern. Specifically, students find analyzing and evaluating easier, while creating has a higher level of difficulty. According to Anderson et al. (2001), in the Revised Bloom Taxonomy, analyzing (C4) and evaluating (C5) skills still involve more systematic processing of information, while creating (C6) requires students to generate new ideas that require a combination of more complex conceptual understanding. It explains why students have more difficulty in the creating category compared to the previous two categories (Fatmawati et al., 2021; Yanti et al., 2023). Zakwandi et al. (2024) research using the Rasch Model indicates that students tend to answer items with a high level of difficulty less correctly, especially when it comes to synthesis skills and creativity.

Furthermore, a study by Brookhart (2010) mentioned that the ability to create requires deep concept mastery, which is often not optimally developed if the learning approach is less oriented towards exploration and open-ended problem-solving. Although the creating aspect has a higher level of difficulty, several other factors can affect the distribution pattern of students' HOTS (Engelhard & Wang, 2021; Ibrahim et al., 2024; Ningsih & Kamaludin, 2023). For example, students may be unfamiliar with problem types that demand creativity or are limited in open-ended problem-solving experiences. Moreover, the memorization-oriented learning approach in schools may hinder the optimal development of HOTS.

Differences in the distribution of students' HOTS based on school origin were found, with some schools exhibiting a higher proportion of students with HOTS than others. Factors that influence student outcomes in HOTS can come from learning quality, school facilities, teachers' pedagogical approaches, and students' backgrounds. The analysis shows that School A and School B have a higher percentage of students with high HOTS than the other schools (4.85 and 5.45%, respectively), while Schools C, D, and E have more students in the low HOTS category (23.03, 15.76, and 17.58%). There are differences in HOTS distribution between schools, which various external and internal factors may influence. According to Heong et al. (2016) research, differences in HOTS can be caused by variations in teaching strategies and learning approaches used in schools. Schools that apply inquiry-based and problem-based learning (PBL) approaches tend to produce students with higher HOTS skills than schools that still use conventional methods based on memorization (Irdalisa et al., 2024; Santosa et al., 2024; Waluyo & Ridlo, 2025).

Furthermore, a conducive learning environment with access to diverse learning resources can contribute to the improvement of students' HOTS. A study by Andrich & Marais (2019) using the Rasch Model showed that schools with teachers who are more accustomed to applying HOTS-based questions in learning evaluation tend to have students with better HOTS levels. Moreover, research by Leou et al. (2006) confirms that support from schools in the form of teacher training and provision of HOTS learning tools has a significant impact on the development of students' higher-order thinking skills. Although there are differences in HOTS distribution between schools, individual factors such as learning motivation, students' learning styles, and support from parents in HOTS achievement (Fathonah et al., 2025; Nesbitt-Hawes, 2005; Satriya & Atun, 2024). Students who have a supportive family environment and are accustomed to problem-solving and idea exploration from an early age can demonstrate high HOTS despite coming from schools with limited resources. Moreover, differences in the level of difficulty of the questions and students' perceptions of HOTS can also affect the results of the distribution analysis.

Analysis using the Rasch Model showed that students' response patterns to questions with different levels of difficulty varied, reflecting different HOTS among students. Questions with higher item difficulty tended to be answered only by students with better HOTS ability, while most students could answer questions with low to moderate item difficulty. The results of the item difficulty analysis show that items with the aspects of "differentiating" and "organizing strategies" have an item difficulty value of -0.67, which indicates that the questions are easier for students to master. In contrast, items with the aspect of "generating" have an item difficulty value of 1.42, which means that it is more difficult for most students. The response pattern shows that students score lower on items with great difficulty, while items with low difficulty are more often answered correctly. According to Andrich & Marais (2019), in the Rasch Model, questions with high difficulty levels can only be answered correctly by students with commensurate or higher ability levels, while almost all students can answer questions with low difficulty levels. It is in line with Smith (2003), which shows that the distribution of response patterns on Rasch-based HOTS tests reflects the extent to which students have analytical, evaluative, and creative thinking skills. Research by Zakwandi et al. (2024) shows that in HOTS-based assessments, students with good critical thinking skills tend to have consistent response patterns on questions with various levels of difficulty, while students with low thinking skills tend to show inconsistencies, especially on questions with high difficulty levels. In addition, the application of the Rasch Model in educational evaluation has proven to be able to identify student response patterns more objectively than the classical method (CTT - Classical Test Theory). Although the Rasch Model provides a more accurate overview of students' response patterns than traditional methods, there are external factors such as students' motivation, test anxiety, and previous experience that can influence their response patterns (Bond et al., 2021). Moreover, technical factors such as understanding the form of the two-tier test questions can also affect the results, especially for students who are not familiar with the format (Treagust, 1988). Thus, the results of the evaluation of students' response patterns based on the Rasch Model show that the level of question difficulty has a significant effect on students' response patterns, with high-difficulty questions being answered more by students with better HOTS. However, students' internal and external factors can also play a role in the variation of responses, so a more holistic approach is needed in assessing students' HOTS skills, including guidance and teaching strategies that are more supportive of HOTS development.

Applicability of Rasch Model in HOTS Evaluation

The two-tier test is an effective assessment instrument in measuring HOTS in science learning because it can identify not only students' final answers but also the reasons behind their choices, thus providing a more accurate picture of students' conceptual understanding and level of critical thinking. According to the results of the validity test analysis using the Rasch Model, most of the items in the Two-Tier Test demonstrate fit with the model, with MNSQ infit and outfit values in the range of 0.5 to 1.5, indicating that the questions are able to measure students' abilities well. Furthermore, the Person Reliability value of 0.93 and Item Reliability of 0.83 indicate that this instrument has a high level of reliability in assessing students' HOTS. The distribution of students' HOTS shows that the majority of students have a low to moderate level of thinking, which indicates that the two-tier test-based assessment can reveal students' difficulties in developing higher-order thinking skills. According to Treagust (1988), the Two-Tier Test is effective in measuring conceptual understanding and detecting students' misconceptions because it requires them to give reasons for the answers chosen. Research by Istiyono (2016), also indicated that the two-tier test is superior to conventional tests in measuring HOTS because it provides more in-depth information about students' thinking patterns, including analysis, evaluation, and creation skills in accordance with the Revised Bloom Taxonomy. Potvin et al. (2015) research found that two-tier tests can enhance the quality of HOTS assessment in science subjects because they help teachers identify the extent to which students really understand concepts and how they apply logical thinking in solving problems.

Furthermore, Zakwandi et al. (2024) stated that the two-tier test is more valid and reliable than the regular multiple-choice test in evaluating HOTS because it can distinguish between students who really understand the concept and students who only guess the answer. Although effective in measuring HOTS, two-tier tests also have challenges, such as difficulties in compiling quality items and taking longer time to analyze student answers (Oktadila et al., 2025; Widiyatmoko & Shimizu, 2018). Several studies have mentioned that students who are not familiar with the two-tier test format have difficulty answering the reasoning section so that the assessment results can be affected by non-cognitive factors, such as reading skills and language comprehension (Cari et al., 2020; Potvin et al., 2015). Therefore, the two-tier test is an effective measurement tool in assessing HOTS in science learning because it can provide a more comprehensive overview of students' conceptual understanding and identify misconceptions more accurately than conventional tests.

Limitations and Suggestions for Future Research

The limited sample coverage in the Purbolinggo sub-district may affect the generalizability of the research findings related to students' HOTS in science learning. It may limit the conclusions that can be applied more broadly to student populations in other areas with different educational conditions. Therefore, further research with wider geographical coverage is needed to obtain a more comprehensive picture of the effectiveness of HOTS assessment in various school conditions and learning environments. Potential bias in HOTS measurement can arise due to various external factors, such as learning methods used by teachers and students' motivation level in learning. Learning methods that do not emphasize critical thinking and problem-solving skills can cause HOTS assessment results not accurately reflect students' cognitive abilities. Therefore, in interpreting the assessment results, it is necessary to control these external factors, such as by combining the assessment results with observations of the learning process or additional instruments that measure aspects of student motivation and engagement. The development of HOTS assessments can be enhanced by combining the Rasch Model with other assessment approaches to obtain a more comprehensive validation. One method that can be used is confirmatory analysis with Structural Equation Modeling (SEM), which allows testing the relationship between HOTS indicators and other factors that influence assessment results.

Conclusion

The study concludes that students' higher-order thinking skills (HOTS) on the topic of Vibration and Waves remain low, with the greatest difficulty found in the creating (C6) aspect, which involves generating original solutions or perspectives. This may be attributed to the limited emphasis on open-ended problem-solving in typical classroom instruction. In contrast, students performed relatively better on analyzing (C4) and evaluating (C5), possibly because these skills align more closely with the structured and procedural nature of science learning in schools. The Two-Tier Test instrument used in this study demonstrated strong psychometric properties, suggesting its effectiveness in capturing varying levels of students' HOTS. Differences in student performance across schools highlight the role of external factors such as instructional quality and teaching methods. To improve HOTS, especially in creating, educators are encouraged to integrate project-based learning, inquiry-based experiments, or design challenges that allow students to invent and evaluate multiple solutions. Additionally, teacher professional development should focus on designing HOTS-aligned assessments and facilitating reflective classroom discussions to deepen students' critical thinking.

Acknowledgement

All author would like to thank to all parties who has supported this research, especially The Research and Innovation Institute (Lembaga Riset dan Inovasi) of Universitas Muhammadiyah Surakarta for the funding.

Author Contributions

Formal analysis, writing—review and editing, A.I.K.; Conceptualization, methodology, investigation, writing—original draft preparation, H.A.; validation, review and editing, L.A.; supervision, review and editing, R.A.; supervision, review and editing, M.W. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Research and Innovation Institute (Lembaga Riset dan Inovasi) of Universitas Muhammadiyah Surakarta under contract number 95.5/A.3-III/LRI/IV/2025. The APC was also funded by the same institution.

Conflicts of Interest

The authors declare no conflict of interest.

References

Affandy, H., Nugraha, D. A., Pratiwi, S. N., & Cari, C. (2021). Calibration for Instrument Argumentation Skills on the Subject of Fluid Statics Using Item Response Theory. *Journal of Physics: Conference Series*, *1842*(1), 1–10. https://doi.org/10.1088/1742-6596/1842/1/012032

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer Nature.

Bond, T. (2015). *Applying the Rasch model: Fundamental Measurement in the Human Sciences* (3rd Ed). Routledge.

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th ed.). Routledge.

Brookhart, S. M. (2010). *How to Assess Higher-Order Thinking Skills in Your Classroom*. ASCD.

Cari, C., Pratiwi, S. N., Affandy, H., & Nugraha, D. A. (2020). Investigation of Undergraduate Student Concept Understanding on Hydrostatic Pressure Using Two-Tier Test. *Journal of Physics: Conference Series*, *1511*(1). https://doi.org/10.1088/1742-6596/1511/1/012085

Engelhard, G., & Wang, J. (2021). *Rasch Models for Solving Measurement Problems: Invariant Measurement in the*

Social Sciences. SAGE Publications.

Fathonah, S., Ahmadi, F., Arsyad, A. A., & Rahman, S. (2025). Problem-Based Learning Model Assisted by Interactive Media to Improve Students Higher Order Thinking Skills (HOTS). *Jurnal Penelitian Pendidikan IPA*, *11*(1), 1234–1243. https://doi.org/10.29303/jppipa.v11i1.6649

Fatmawati, B., Wazni, M. K., & Husnawati, N. (2021). The Study of Worksheets Based on Creative Problem Solving for Biology Subjects. *Jurnal Penelitian Pendidikan IPA*, *7*(4), 701–706. https://doi.org/10.29303/jppipa.v7i4.831

Fensham, P. J., & Bellocchi, A. (2013). Higher Order Thinking in Chemistry Curriculum and Its Assessment. *Thinking Skills and Creativity*, *10*, 250–264. https://doi.org/10.1016/j.tsc.2013.06.003

Heong, Y. M., Sern, L. C., Kiong, T. T., & Mohamad, M. M. B. (2016). The Role of Higher Order Thinking Skills in Green Skill Development. *EDP Sciences*, *70*(05001), 1–5. https://doi.org/10.1051/matecconf/20167005001

Hidayatullah, A. R., Yamtinah, S., & Masykuri, M. (2022). Development of A Two-Tier Multiple-Choice Instrument Based on Higher Order Thinking Skills (HOTS) on Acids, Bases, and Salts. *Jurnal Penelitian Pendidikan IPA*, *8*(2), 932–938. https://doi.org/10.29303/jppipa.v8i2.1423

Ibrahim, N. M., Sanjaya, Y., & Nurjhani, M. (2024). Effectiveness of Biology Learning to Improve Digital Literacy and Higher Order Thinking Skills on the Concept of Digestive System. *Jurnal Penelitian Pendidikan IPA*, *10*(9), 7131–7137. https://doi.org/10.29303/jppipa.v10i9.5018

Irdalisa, I., Akbar, B., Fuadi, T. M., Maesaroh, M., & Kartikawati, E. (2024). Ricosre Model with Question Formulation Technique (QFT): Enhancing Students' Higher Order Thinking Skills (HOTS) and Science Literacy. *Jurnal Penelitian Pendidikan IPA*, *10*(3), 1175–1178. https://doi.org/10.29303/jppipa.v10i3.6764

Istiyono, E. (2016). The Application of GPCM on MMC Test as a Fair Alternative Assessment Model in Physics Learning. *Proceeding of 3rd International Conference on Research, Implementation and Education of Mathematics and Science*, 25–30.

Istiyono, E., Dwandaru, W. B., & Rahayu, F. (2018). Pengembangan Tes Creative Thinking Skills Fisika SMA (PhysCreTHOTS) Berdasarkan Teori Tes Modern. *Cakrawala Pendidikan*, *37*(2), 190–200. Retrieved from https://journal.uny.ac.id/index.php/cp/article/download/19233/11107

Istiyono, E., Mustakim, S. S., Widihastuti, W., Suranto, S., & Mukti, T. S. (2019). Measurement of Physics Problem-Solving Skills in Female and Male Students by PhysTeProSS. *Jurnal Pendidikan IPA Indonesia*, *8*(2), 170–176. https://doi.org/10.15294/jpii.v8i2.17640

Jansen, T., & Meoller, J. (2022). Teacher Judgments in School Exams: Influences of Students' Lower Order Thinking Skills on the Assessment of Students' Higher Order Thinking Skills. *Teaching and Teacher Education*, *111*, 1–10. https://doi.org/10.1016/j.tate.2021.103616

Kaltakci, D., Eryilmaz, A., & McDermott, L. C. (2016). Identifying Pre-Service Physics Teachers' Misconceptions and Conceptual Difficulties About Geometrical Optics. *European Journal of Physics*, *37*(4), 045705. https://doi.org/10.1088/0143-0807/37/4/045705

Khaeruddin, K., Indarwati, S., Sukmawati, S., Hasriana, H., & Afifah, F. (2023). An Analysis of Students' Higher Order Thinking Skills Through the Project-Based Learning Model on Science Subject. *Jurnal Pendidikan Fisika Indonesia*, *19*(1), 47–54. https://doi.org/10.15294/jpfi.v19i1.34259

Kwangmuang, P., Jarutkamolpong, S., Sangboonraung, W., & Daungtod, S. (2021). The Development of Learning Innovation to Enhance Higher Order Thinking Skills for Students in Thailand Junior High Schools. *Heliyon*, *7*(6), e07309. https://doi.org/10.1016/j.heliyon.2021.e07309

Leou, M., Abder, P., Riordan, M., & Zoller, U. (2006). Using "HOCS-Centered Learning" as a Pathway to Promote Science Teachers' Metacognitive Development. *Research in Science Education*, *36*(1–2), 69–84. https://doi.org/10.1007/s11165-005-3916-9

Lewis, A., & Smith, D. (1993). Defining Higher Order Thinking. *Theory Into Practice*, *32*(3), 131–137. https://doi.org/10.1080/00405849309543588

Linacre, J. M. (2011). *A User's Guide to Winsteps Rasch Model Computer Programs*. Chicago

Negara, A. H. S., Waston, W., Hidayat, S., & Mulkhan, A. M. (2024). Development of Religious Character to Improve the Effectiveness of Teacher and Student Communication. *Revista de Gestao Social e Ambiental*, *18*(6), 1–26. https://doi.org/10.24857/rgsa.v18n6-037

Nesbitt-Hawes, P. J. (2005). *Higher Order Thinking Skills in a Science Classroom Computer Simulation* [Queensland University of Technology]. https://doi.org/10.1515/9783050062365

Ningsih, N. R., & Kamaludin, A. (2023). Development of Higher Order Thinking Skills-Based Assessment Instrument on Acid-Base Materials in High School. *Jurnal Penelitian Pendidikan IPA*, *9*(1), 13–19. https://doi.org/10.29303/jppipa.v9i1.1457

Oktadila, S. S., Nasbey, H., & Jaya, I. (2025). Enhancing Higher-Order Thinking Skills in Elementary Science Learning Using the RADEC Model. *Jurnal Penelitian*

Pendidikan IPA, *11*(3), 573–579. https://doi.org/10.29303/jppipa.v11i3.10398

PISA. (2023). *Programme for International Student Assessment*. Retrieved from https://www.oecd.org/

Potvin, P., Skelling-Desmeules, Y., & Sy, O. (2015). Exploring Secondary Students' Conceptions About Fire Using a Two-Tier, True/False, Easy-to-Use Diagnostic Test. *Journal of Education in Science, Environment and Health*, *1*(2), 63. https://doi.org/10.21891/jeseh.99647

Ramadan, Z. H., Anggriani, M. D., & Dafit, F. (2023). Development of 3-Dimensional Cartoon Animation Videos to Improve Higher Order Thinking Skills of Elementary School Students. *Jurnal Penelitian Pendidikan IPA*, *9*(11), 10506–40516. https://doi.org/10.29303/jppipa.v9i11.5564

Santosa, T. A., Angreni, S., Sari, R. T., Festiyed, F., Yerimadesi, Y., Ahda, Y., Alberida, H., & Arsih, F. (2024). Effectiveness of Higher Order Thinking Skills-based Test Instruments in Science Learning in Indonesia: A Meta-analysis. *Jurnal Penelitian Pendidikan IPA*, *10*(5), 242–249. https://doi.org/10.29303/jppipa.v10i5.6654

Satriya, M. A., & Atun, S. (2024). The Effect of Argument Driven Inquiry Learning Models on Scientific Argumentation Skills and Higher Order Students on The Topics of Acid Base. *Jurnal Penelitian Pendidikan IPA*, *10*(5), 2663–2673. https://doi.org/10.29303/jppipa.v10i5.6834

Seibert, S. A. (2021). Problem-Based Learning: A Strategy to Foster Generation Z's Critical Thinking and Perseverance. *Teaching and Learning in Nursing*, *16*(1), 85–88. https://doi.org/10.1016/j.teln.2020.09.002

Smith, R. M. (2003). *Rasch Measurement Models: Interpreting WINSTEPS/BIGSTEPS and FACETS Output*. JAM Press.

Susantini, E., Isnawati, I., & Raharjo, R. (2022). HOTS-Link Mobile Learning Application: Enabling Biology Pre-Service Teachers to Devise HOTS-Based Lesson Plans. *Journal of Science Education and Technology*, *31*, 783–794. https://doi.org/10.1007/s10956-022-09993-w

Treagust, D. F. (1988). Development and Use of Diagnostic Tests to Evaluate Students' Misconceptions in Science. *International Journal of Science Education*, *10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Waluyo, J., & Ridlo, Z. R. (2025). Implementation of Science Module Based on Microcontroller to Improve Students' Computational Thinking Skills in Earth Science Course Thinking. *Jurnal Penelitian Pendidikan IPA*, *11*(2), 752–760. https://doi.org/10.29303/jppipa.v11i2.10348

Widiyatmoko, A., & Shimizu, K. (2018). The Development of Two-Tier Multiple Choice Test to Assess Students' Conceptual Understanding About Light and Optical Instruments. *Jurnal Pendidikan IPA Indonesia*, *7*(4), 491–501. https://doi.org/10.15294/jpii.v7i4.16591

Yanti, N. L. I. M., Redhana, W., & Suastra, W. (2023). Multiple Scaffolding STEAM Project-Based Learning Model In Science Learning. *Jurnal Penelitian Pendidikan IPA*, *9*(8), 6493–6502. https://doi.org/10.29303/jppipa.v9i8.4470

Zakwandi, R., Istiyono, E., & Dwandaru, W. S. B. (2024). A Two-Tier Computerized Adaptive Test to Measure Student Computational Thinking Skills. *Education and Information Technologies*, *29*(7), 8579–8608. https://doi.org/10.1007/s10639-023-12093-w