REID (Research and Evaluation in Education), 11(1), 2025, 1-16
Available online at: http://journal.uny.ac.id/index.php/reid

Developing and analyzing items of a physics conceptual understanding test on wave topics for high school students using the Rasch Model

Fauziah Rasyid*; Edi Istiyono; Cahya Widya Gunawan
Universitas Negeri Yogyakarta, Indonesia
*Corresponding Author. E-mail: fauziahrasyid.2023@student.uny.ac.id

ARTICLE INFO

Article History
Submitted:
June 28, 2024
Revised:
November 1, 2024
Accepted:
November 23, 2024

Keywords
conceptual understanding; instrument development; item analysis; Rasch Model

Scan Me:

ABSTRACT

This study aims to develop, validate, and analyze test items for assessing the understanding of mechanical wave concepts among high school students. The test development process followed the Mardapi instrument development model, which includes: (1) constructing test specifications, (2) writing test items, (3) reviewing test items, (4) piloting the test, and (5) analyzing the items. The developed instrument consists of 12 multiple-choice items, covering three aspects of conceptual understanding: translation, interpretation, and interpolation. Content validity was assessed by three validators, and the results were analyzed using the Aiken V method. The instrument was then administered to 257 high school students in South Sulawesi Province. The results were analyzed using Item Response Theory (IRT) with the Rasch model through the Quest program. Item analysis included item fit estimation, reliability, and item difficulty. The content validity test results indicate that the instrument is valid. All items fit the Rasch model, with a reliability coefficient of 0.95, categorized as high reliability. Item difficulty analysis revealed that 8.3% of items were categorized as easy, 8.3% as difficult, and 83.3% as moderate. Overall, the results indicate that the test instrument is of good quality and can be used to assess high school students' understanding of mechanical wave concepts.

To cite this article (in APA style):
Rasyid, F., Istiyono, E., & Gunawan, C. W. Developing and analyzing items of a physics conceptual understanding test on wave topics for high school students using the Rasch Model. *REID (Research and Evaluation in Education)*, *11*(1), 1-16. https://doi.org/10.21831/reid.v11i1.75575

INTRODUCTION

The most fundamental ability to master 21st-century skills is conceptual understanding. One effort in implementing learning to enhance these 21st-century skills is through productive learning practices that enrich knowledge, aiming to achieve conceptual understanding (Darling-Hammond et al., 2020). Students' conceptual understanding of physics content is crucial for fostering their higher-order thinking skills, which enable them to solve everyday problems creatively (Putranta & Supahar, 2019). Understanding concepts in physics plays a significant role in learning (Putri et al., 2020). Furthermore, understanding is a cognitive aspect that significantly contributes to the success of student learning (Sartika, 2018). A poor grasp of concepts can lead to numerous issues, one of which is the occurrence of misconceptions (Kola, 2017). Therefore, physics learning aimed at improving students' understanding of physics concepts is essential.

The understanding of physics concepts is not only important but also presents a challenge for students. Physics content often poses challenges due to its complexity and the abstract nature of its concepts (Pals et al., 2023). One topic in physics that presents a challenge is the topic of waves (Sharma et al., 2023). A comprehensive understanding of mechanical waves is essential for students to succeed in mastering various advanced physics topics (F. Kurniawan et al., 2023; Xie

et al., 2021), as it enables them to develop a deeper understanding of more abstract wave phenomena, such as electromagnetic waves (Goodhew et al., 2019). However, the concept of mechanical waves has been identified as a difficult topic, and students face conceptual difficulties related to wave topics at various educational levels (Kanyesigye et al., 2022). Hence, efforts are needed in physics learning to facilitate students' understanding of mechanical wave concepts.

Effective physics learning that supports students' conceptual understanding can be influenced by various components, one of which is assessment. Assessment has a significant impact on improving students' abilities in physics learning (Chen et al., 2018; Darling-Hammond et al., 2020). A strategy that can help students understand physics material well is by administering tests to measure their level of understanding (Larasati et al., 2020). Thus, a high-quality instrument is necessary as a tool to measure concept understanding while supporting the effectiveness of physics learning. Furthermore, well-developed test instruments can also serve as valuable tools in educational research to assess the effectiveness of teaching methods and curricula.

Developing a high-quality test instrument requires an appropriate approach to ensure its validity, reliability, and objectivity. The Rasch model, as part of Item Response Theory (IRT), helps improve the accuracy of instrument development, ensures its quality, and accurately computes respondents' performances (Boone, 2016). The Rasch model provides objective measurements based on student ability, which are unaffected by item difficulty levels in the assessment task (Asriadi & Hadi, 2021). The Rasch model plays a key role in developing instruments that meet the principles of scientific measurement, including testing invariance and scale precision (Bond et al., 2020). Rasch modeling is a robust analytical approach for assessing item performance, detecting shifts in difficulty over time, and enabling comparisons of students who participated in assessments at different times or locations (Hope et al., 2024). Rasch is also widely used to validate educational tests to ensure they are free from bias and align with the abilities measured in diverse populations (Bond et al., 2020). Therefore, the Rasch model is an effective approach in developing high-quality test instruments.

Previous studies have used the Rasch model in test development for wave topics. Mešić et al. (2019) developed a conceptual understanding test for the wave optics topic for university physics students using the Rasch model. Similarly, Balta et al. (2022) developed a test for high school students on wave optics concepts using Rasch analysis. The study by A. Kurniawan et al. (2024) analyzed the quality of test items in an instrument designed to assess high school students' conceptual understanding of the electromagnetic wave topic. However, studies focusing on developing tests for conceptual understanding of mechanical wave topics for high school students using the Rasch model are still limited. Thus, the researchers aim to develop a test instrument to assess conceptual understanding of the topic of mechanical waves. This study aims to develop, validate, and analyze test items to investigate high school students' conceptual understanding of mechanical wave concepts based on the Rasch model in Item Response Theory (IRT).

METHOD

Research Design

This study employs a development research method with a quantitative approach. The research aims to develop, validate, and analyze a concept understanding test instrument on the topic of mechanical waves for high school students. The development of the test instrument follows Mardapi's (2008) instrument development procedure: (1) developing test specifications, (2) constructing test items, (3) reviewing test items, (4) conducting preliminary testing, (5) analyzing items, (6) revising test items, (7) assembling the test, (8) implementing the test, and (9) interpreting test results (Mardapi, 2008). This study concludes its focus on the item analysis stage.

Data Collection and Sample Size

Data collection was carried out through the implementation of the concept understanding test on mechanical wave material using a Google Form. The instrument consisted of 12 items, tested on a sample of 257 students from grade XI in 11 high schools in South Sulawesi. The minimum sample size requirement for Rasch model analysis, according to Wright and Stone (1979), is 200 respondents. Şahin and Anıl (2017) recommend a minimum of 150 respondents for the Rasch model, while Mešić et al. (2019) suggest at least 100 samples to ensure sufficiently stable parameter estimation in Rasch-based concept understanding tests. Therefore, the sample size of 257 respondents in this study is adequate.

Data Analysis

The analysis of the test instrument was conducted to assess the quality of items designed to measure high school students' conceptual understanding of mechanical waves. The developed instrument underwent a rigorous content validity evaluation, employing Aiken's V method (Aiken, 1980) to determine its content validity.

Content Validity

The content validity of the test instrument was assessed by three expert validators using a structured validation framework. The validation instrument consisted of 15 evaluation statements grouped into three core criteria: content accuracy, structural integrity, and linguistic clarity. The assessment results encompassed numerical validity scores and constructive feedback from the validators, aimed at enhancing the quality of the test items. Recommendations from the validators were systematically incorporated to refine and improve the instrument. The validation scores were analyzed using Aiken's V formula, providing a content validity coefficient that reflects the instrument's alignment with theoretical and practical standards (Aiken, 1980), as presented in Formula (1). Meanwhile, the content validity coefficient obtained is categorized based on Table 1 (Istiyono, 2020).

$$V = \frac{\sum s}{n(c-1)} \quad \text{..... (1)}$$

with:
V : content validity coefficient
s : $r - l_0$
l_0 : lowest validity score assigned
c : highest validity score possible
r : score provided by a validator

Table 1. Content Validity Categories

Validity Coefficient	Category
$V < 0.4$	Low validity
$0.4 < V \leq 0.8$	Moderate validity
$V > 0.8$	High validity

Rasch Model

The Rasch model is currently considered the most reliable approach to the fundamental principles of measurement in the humanities (Bond et al., 2020). One simple variant of the Rasch model is the dichotomous Rasch model, where the relationship between an individual's ability and the difficulty of an item is explained through probabilities (Boone et al., 2014). The main advantage of the Rasch model lies in its specific objectivity feature, where differences in item difficulty estimation are independent of the sample (Szabó, 2008). A key characteristic of the

Rasch model is its ability to describe the relationship between an individual's ability and the difficulty level of an item, where the probability of success in answering an item is determined by the difference between the individual's ability and the item's difficulty (Bond et al., 2020). The benefits of the Rasch model include: (a) evaluating whether the test items fit and identifying potential item biases; (b) item calibration that is not influenced by sample ability; (c) using the standard error of calibration to assess the precision of each item; and (d) estimating item difficulty from various samples and converting them to a common scale (Wright, 1977). With its various advantages, the Rasch model is a robust and objective tool for determining the quality of test items.

The Rasch model is used to analyze the trial results of the instrument with 257 respondents. The instrument is considered to be of good quality if it meets the criteria for evaluating item assessment, which includes the stages of: (1) estimation of item suitability; (2) estimation of difficulty level; (3) estimation of item fit; and (4) estimation of reliability (Hanna & Retnawati, 2022). The first stage involves estimating item suitability based on the Infit Mean Square value, and the third stage estimates item fit based on outfit t. Infit t and outfit t indicate the instrument's validity based on its fit with the Rasch model (I. Azizah & Supahar, 2023). Therefore, this study analyzes test items to evaluate their quality based on (1) Validity (relevant to the Rasch model), (2) Reliability, and (3) Item difficulty. The Rasch model analysis is performed using the Quest program. The primary element of the Quest program is Item Response Theory (IRT) adjusted to the Rasch model (Habibi et al., 2019). The results of the instrument trial are used to analyze item quality using the Rasch model, including item fit, reliability, and item difficulty through the Quest program.

Item Fit

Fit statistics generally focus on two aspects of item fit, namely infit and outfit (Bond et al., 2020). Items will be considered valid if they align with the Rasch model, with the INFIT MNSQ value falling within the range of 0.77–1.33 and the OUTFIT t value ≤ 2 with a probability of 0.5 (Dewi et al., 2023; Lafifa & Dadan, 2024). The range for infit or outfit MNSQ of $0.7 < \text{MNSQ} < 1.3$ is used as a guideline to assess item fit in Rasch analysis (I. Azizah & Supahar, 2023; Bond et al., 2020). This range is considered adequate to detect significant item misfit, though its sensitivity to large violations of the model is still debated (Bond et al., 2020). In this study, the INFIT MNSQ range and outfit t values from the Quest output are used to consider item fit.

Reliability

To test the reliability of this developed test instrument, the Quest program was used. The output from Quest provides two types of reliability: reliability of items and case estimates (Rahim et al., 2023). The reliability results of the test instrument can be categorized based on the reliability coefficient value, as shown in Table 2 (Istiyono, 2020).

Table 2. Reliability Categories

Reliability Coefficient	Category
0.80 – 1.00	Very High
0.60 – 0.80	High
0.40 – 0.60	Medium
0.20 – 0.40	Low
-1.00 – 0.20	Very Low

Item Difficulty

Based on the Quest output, the difficulty index (parameter b) of each item from the developed test instrument is obtained by examining the Thresholds (THRSH) (Setyawarno, 2017). A question item is considered good if its difficulty index (b) ranges from -2 to +2, when

the ability score is transformed so that it has an average of 0 and a standard deviation of 1 (Sumaryanta, 2021). If the b value approaches -2, the item's difficulty index is very low, whereas if the *b* value approaches +2, the item's difficulty index is very high. The criteria for item difficulty are presented in Table 3 (Baker, 2001).

Table 3. Category of Item Difficulty

Difficulty Index	Category
$b > 2$	Very difficult
$1 < b < 2$	Difficult
$-1 \leq b \leq 1$	Medium
$-1 > b \geq -2$	Easy
$b \geq -2$	Very easy

FINDINGS AND DISCUSSION

Findings

Instrument Development

The first step in instrument development is constructing test specifications. This process involves defining the objectives of measurement, the scope of content, and the competencies to be assessed. The test specifications are systematically organized into a matrix or blueprint. The objective of the developed test instrument is to assess cognitive learning outcomes in physics for 11th-grade high school students. The competency targeted for measurement is conceptual understanding. The content scope includes topics on mechanical waves aligned with the Indonesia Merdeka Curriculum Phase F, covering quantities of mechanical waves, travelling waves, and standing waves. These test specifications are systematically organized into a matrix or blueprint to ensure alignment with the intended goals. The development of the instrument adheres to the matrix (blueprint) that has been systematically designed. Table 4 provides a presentation of the developed question matrix, illustrating its alignment with the specified objectives and competencies.

Table 4. Instrument Blueprint

Aspects	Sub-aspect	Mechanical Wave Material		
		Transverse and longitudinal waves	Traveling wave	Stationary wave
Concept Understanding	Translation	1, 2	6	9
	Interpretation	3, 4	7	10
	Extrapolation	-	5, 8	11, 12

The second development step is constructing test items. At this stage, a matrix (blueprint) is prepared, which is developed into a question grid. Based on the grid, the items were then developed into 12 multiple-choice items.

The third development step is reviewing test items. The test instruments that have been developed are reviewed or tested for content validity by three validators. The developed test instrument was reviewed or tested for content validity by three validators. The validator's assessment results were calculated using the Aiken V content validity formula. The content validity results shown in Table 5, that the scores for each item on the instrument were consistently high, indicating strong agreement among the validators. This suggests that the instrument is effective and reliable in measuring the intended concepts, providing a robust assessment tool. An Aiken V value approaching 1 indicates high content validity, whereas lower values suggest poor validity.

Table 5. Content Validity

Item Number	Validity Coefficients	Category
1	1.00	High validity
2	0.89	High validity
3	1.00	High validity
4	1.00	High validity
5	0.89	High validity
6	0.89	High validity
7	1.00	High validity
8	0.89	High validity
9	1.00	High validity
10	1.00	High validity
11	1.00	High validity
12	1.00	High validity

Table 5 presents all items in the test instrument for understanding mechanical wave concepts are declared valid. The validity values obtained are close to 1, and some even have a value of 1, indicating that the instrument has high content validity.

The fourth step of instrument development is conducting test trials. The assembled test was administered to 257 eleventh-grade high school students to assess their conceptual understanding of the topic of mechanical waves. Their responses submitted via Google Forms were exported in a spreadsheet format. These responses were then converted into dichotomous items, with correct answers scored as 1 and incorrect answers scored as 0. The responses were subsequently analyzed for item fit, reliability, and item difficulty levels based on the Rasch model.

Validity of Instrument

Validity (fit with the Rasch model) of the instrument can be obtained from Infit Meansquare and Outfit t value. The Quest output that displays the Infit Meansquare distribution is illustrated in Figure 1.

[Test for the Conceptual Understanding of Dichotomous Mechanical Waves (12 items of columns 7-18)]

```
Tes Pemahaman Konsep Gelombangmekanik dikotomus (12 item kolom 7-18)
-------------------------------------------------------------------------------------------------------------
Item Fit                                                                                    16/ 6/24 10:58
all on all (N = 257 L = 12 Probability Level= .50)
-------------------------------------------------------------------------------------------------------------
INFIT
 MNSQ          .56       .63       .71       .83      1.00      1.20      1.40      1.60      1.80
----------------+---------+---------+---------+---------+---------+---------+---------+---------+---------
  1 item 1                              .               *               .
  2 item 2                              .            *  |               .
  3 item 3                              .            *  |               .
  4 item 4                              .         *     |               .
  5 item 5                              .       *       |               .
  6 item 6                              .               |       *       .
  7 item 7                              .               |  *            .
  8 item 8                              .               |  *            .
  9 item 9                              .               *               .
 10 item 10                             .               *|              .
 11 item 11                             .               |       *       .
 12 item 12                             .               |   *           .
=============================================================================================================
```

Figure 1. Item Fit Quest Output

An item is considered to fit the Rasch model when it has an infit mean square (MNSQ) value ranging from 0.7 to 1.3 and an outfit t-value less than 0.2. Table 6 presents information on Infit Meansquare, Outfit t and categorization of the validity of each item of the mechanical wave topic concept understanding test instrument.

Table 6. Item Validity Based on Fit with the Rasch Model

Item Number	MNSQ Infit	Outfit t	Interpretation
1	1.00	0.7	Valid
2	0.94	-0.7	Valid
3	0.95	-0.3	Valid
4	0.87	-1.9	Valid
5	0.86	-1.3	Valid
6	1.15	0.7	Valid
7	1.03	0.3	Valid
8	1.04	0.4	Valid
9	1.01	-0.3	Valid
10	0.99	-0.2	Valid
11	1.16	1.7	Valid
12	1.06	0.8	Valid

Table 6 indicates that all items used fall within the acceptable range according to the Rasch model criteria based on INFIT MNSQ and INFIT values. All items have met the validity criteria, and therefore, no revisions are necessary. Consequently, the analysis results can be fully utilized to interpret the probability of students' responses in understanding concepts related to mechanical waves. This suggests that the developed items are appropriate and valid for measuring students' conceptual understanding of the mechanical wave topic.

Reliability of Instrument

The test instrument reliability results obtained can be categorized based on the reliability coefficient value, as previously specified in Table 2. The reliability of the instrument, based on the Quest output, includes both item reliability and case reliability, which are displayed in Figure 2. The Quest program provides these reliability estimates to evaluate how well the test items function and how consistent the measurements are across different respondents. Item reliability assesses the stability of item parameters, ensuring that each item measures what it intends to across different samples, while case reliability looks at the consistency of measurements for individual respondents, confirming that the instrument can reliably distinguish between different levels of ability or performance. These reliability indices are crucial for determining the overall quality of the test instrument and ensuring its validity in measuring the intended construct.

```
Summary of item Estimates              Summary of case Estimates
=========================              =========================

Mean                        .00        Mean                      -1.00
SD                          .75        SD                         1.20
SD (adjusted)               .74        SD (adjusted)               .91
Reliability of estimate     .95        Reliability of estimate     .58
```

Figure 2. Reliability Quest Output

Figure 2 shows the results of item reliability and case reliability. The item reliability score is 0.95, which, based on Table 4, falls into the high reliability category. It indicates that the test instrument for conceptual understanding of the wave topic has excellent item quality. Meanwhile, the case reliability score is 0.58, which falls into the moderate reliability category. This value reflects the consistency of respondents in completing the test. Although a case reliability score of 0.58 indicates a reasonable level of consistency, further improvements are necessary to achieve optimal results.

Difficulty Index

The item difficulty index or parameter b represents the difficulty level of a test item. The item difficulty of the developed instrument can be seen in Table 7.

Table 7. Item Difficulty Index

Item Number	b	Category
1	-1.23	Easy
2	-0.21	Medium
3	-0.23	Medium
4	-0.87	Medium
5	0.93	Medium
6	0.19	Medium
7	0,19	Medium
8	-0.52	Medium
9	0.64	Medium
10	-0.39	Medium
11	1.45	Hard
12	0.05	Medium

Table 8. Item Difficulty Distribution and Quality Judgement

Difficulty Categories	Quality	Item Number	Frequency	Percentage
Easy	Good	1	1	8.33%
Medium	Good	2, 3, 4, 5, 6, 7, 8, 9, 10, 12	10	83.33%
Hard	Good	11	1	8.33%

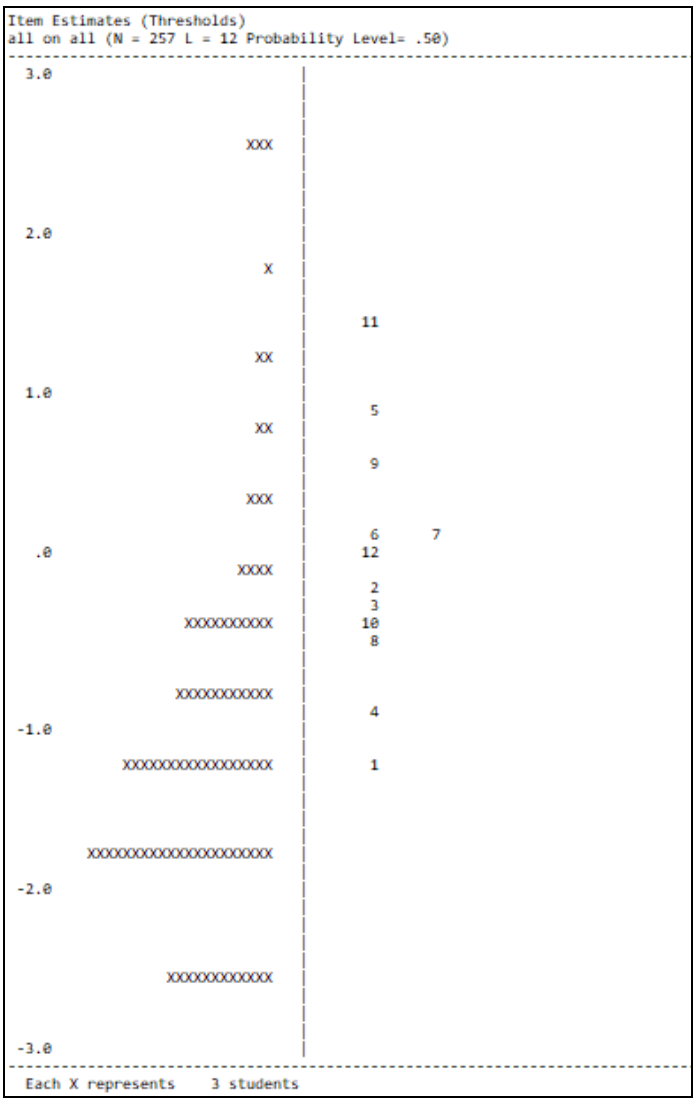

Figure 3. Item Difficulty and Ability Distribution

Table 7 presents parameter b values ranging from -1.23 to 1.45, categorized into easy, medium, and hard items, with the majority of test items classified as having a medium difficulty level. Table 8 illustrates the distribution of items and their quality assessment based on item difficulty levels. All items have difficulty indices within the range of -2 to 2, indicating that all items are of good quality in terms of difficulty level. The developed test exhibits a balanced distribution of difficulty levels, dominated by medium-difficulty items (83.33%), while easy and hard items are equally represented (8.33% each).

Figure 3 is also called item maps or wright maps. The vertical axis shows the difficulty or proficiency scale. The left part of the figure depicts the distribution of examinees' abilities, and the right part shows the location of item difficulty. Figure 3 illustrates the logit range from -3 to +3, where items located near logit -3 are categorized as low-difficulty items, while items around logit +3 are classified as high-difficulty items. Items near logit 0 are considered to have moderate difficulty levels. The easiest item is located around logit -1, whereas the most difficult item is above logit +1. The Wright Map also depicts the distribution of students' ability levels, enabling the identification of students with similar abilities based on identical logit values. Each "X" on the map represents three test takers. As shown in Figure 3, the majority of test takers have abilities below logit 0, which represents the average ability level. Only a small proportion of test takers demonstrate abilities above logit 0, reflecting performance above the average level.

Discussion

Test Specification

The instrument developed based on the development steps outlined by Mardapi (2008) resulted in 12 multiple-choice items designed to measure high school students' conceptual understanding of physics in the topic of mechanical waves. The aspects of conceptual understanding include translation, interpretation, and extrapolation (Halim et al., 2017). The chosen test format is multiple-choice, which is widely used in physics education research literature. For instance, Planinic et al. (2024) employed a multiple-choice test instrument to assess students' understanding of wave optics. Similarly, studies by Mešić et al. (2019) and Balta et al. (2022) developed multiple-choice test instruments to measure conceptual understanding in wave optics. Multiple-choice tests are an economical and practical method for assessing students' conceptual understanding, offering high objectivity and ease of implementation in large groups (Balta et al., 2022). They also assess various cognitive levels and can cover a broad range of material (Istiyono, 2020). In line with this statement, there is research comparing students' multiple-choice test scores with instructor-rated written explanations. The consistency between multiple-choice scores and the instructor's assessment of student explanations suggests that multiple-choice exams effectively measure student understanding in a statistically equivalent manner (Docktor & Mestre, 2014), so the multiple-choice test form can be an adequate instrument to measure concept understanding.

Content Validity

The developed test instrument was validated for its content validity based on expert validator assessments. Content validity aims to identify the relevance and representativeness of the instrument in relation to the evaluated aspects (Retnawati & Wulandari, 2019). Content validity is determined by experts through an evaluation of the extent to which the content aligns with and appropriately represents the operational definition of the construct being measured (Andrich & Marais, 2019). In general, content validity encompasses representativeness, content relevance and technical quality (Yim et al., 2024). Content validity testing can be conducted using an expert validator assessment approach. Content validity through the validator approach involves evaluating the alignment of test items with a specific content domain based on expert judgment (Kurpius & Stafford, 2005). Table 5 shows that most instrument items have a validity coefficient of 1, with the remainder approaching 1, indicating that all items possess strong

content validity. Based on the validators' assessment, the test instrument for conceptual understanding in the topic of mechanical waves for 11th-grade high school students is deemed suitable in terms of content validity.

Item Analysis with the IRT Rasch Model

The IRT approach was developed as a resolution to address the shortcomings found in classical measurement theory. In classical test theory, item characteristics depend on the group of participants taking the test, and ability is measured based on observed performance scores (Shanti et al., 2020). In this study, item analysis was based on IRT with the Rasch model. The item analysis includes item fit, reliability and item difficulty.

Item Fit

The Rasch model analysis produces fit statistics, which provide insights into whether the obtained data accurately reflect individuals' abilities to respond to test items based on their difficulty levels. This evaluation utilizes infit and outfit parameters, derived from mean square values and their standardized counterparts. Items deemed to fit the model indicate that their behavior aligns consistently with the expectations of the Rasch model (N. Azizah et al., 2022). This study calculates the infit and outfit statistics, which are used to estimate the extent to which an item is useful in distinguishing respondents around the mean score (infit) and at the extremes of the distribution (outfit) (Hope et al., 2024).

Table 6 demonstrates that all test items fall within the acceptable range for INFIT MNSQ and INFIT values based on Rasch model criteria, confirming their validity without any items being discarded. These items effectively measure individual abilities in responding to test items based on their difficulty, exhibiting consistency in alignment with the expectations of the Rasch model. Figure 1 further supports this conclusion, showing that all 12 instrument items fall within the acceptable range of 0.77 to 1.30, thus fitting the Rasch model criteria.

Items that fit the Rasch model are inherently valid, meaning they reliably predict individual performance within appropriate contexts (I. Azizah & Supahar, 2023; Andrich & Marais, 2019). This ensures that the developed instrument provides accurate and meaningful insights into students' conceptual understanding of mechanical waves. Consequently, the findings confirm the instrument's suitability for assessing such understanding, offering a robust tool for evaluating and improving physics education.

Reliability

Reliability is a critical factor in evaluating the quality of an instrument's items. A reliable instrument ensures consistent data collection across repeated assessments within similar populations (Andrich & Marais, 2019). The Quest software provides two key reliability metrics: item reliability and case reliability, where case reliability is also referred to as person reliability (Dewi et al., 2023).

Figure 2 reveals an item reliability score of 0.95, which is categorized as very high. The item reliability index indicates how consistently the placement of items measures abilities when the same items are administered to another group of participants with a similar size and comparable behavior (Bond et al., 2020). It indicates strong internal consistency among the items, reflecting the high quality and robustness of the instrument's construction. A high item reliability score confirms that the test items consistently measure the intended construct across respondents.

In contrast, the case reliability score in Figure 2 is 0.58, categorized as moderate. The person reliability index reflects the consistency of participant score rankings when they are administered a similar set of items measuring the same construct (Bond et al., 2020). In other words, high person reliability indicates a clear pattern where some participants consistently achieve higher scores while others score lower, and this pattern is expected to remain stable (Bond et al., 2020). This value highlights that while the instrument performs well in terms of

item-level consistency, it shows limitations in classifying individuals along the measured latent trait. This finding aligns with previous studies, including those conducted by Dewi et al. (2023), A. Kurniawan et al. (2024), and Faradillah and Febriani (2021). While Dewi et al. (2023) suggested that lower case reliability might result from smaller sample sizes, empirical evidence does not consistently support this claim. For instance, these authors observed a case reliability of 0.26 with a sample of 38 respondents. Faradillah and Febriani (2021) reported a case reliability of 0.53 from a sample of 204 respondents, whereas A. Kurniawan et al. (2024) found a case reliability of 0.25 with 298 respondents. These variations suggest that case reliability is influenced more by respondent consistency and engagement than by sample size alone.

Moderate case reliability in the present study may also reflect challenges associated with administering the test online. Limited supervision and suboptimal respondent conditions likely contributed to inconsistencies in responses. The quality of respondent engagement with test items plays a vital role in determining both reliability and validity. Ensuring that respondents are well-prepared, understand the test format, and are presented with questions in a logical sequence, where easier items are not clustered at the end, can improve their engagement and response consistency (Andrich & Marais, 2019).

Although the observed case reliability of 0.58 falls within an acceptable range, improvements are necessary to enhance the instrument's overall reliability. Refining the arrangement of item difficulty levels, ensuring optimal testing conditions, and promoting respondent consistency are crucial steps in achieving this goal. These adjustments would not only improve case reliability but also strengthen the instrument's ability to classify individuals accurately along the latent trait being measured.

Item Difficulty

Item difficulty can be used to assess whether a test item is targeted to the ability level where measurement precision is desired. Table 7 demonstrates that all 12 conceptual understanding test items on the topic of mechanical waves fall within an acceptable difficulty range. One item (8.3%) is categorized as easy, specifically the first item. Ten items (83.3%) are of moderate difficulty, including items 2, 3, 4, 5, 6, 7, 8, 9, 10, and 12. Meanwhile, one item (8.3%) is categorized as difficult, which is item 11.

The arrangement of test items based on difficulty level can significantly impact the quality of respondents' answers. According to Andrich and Marais (2019), placing easier questions at the beginning and more difficult ones toward the end is advisable. Structuring the test from the easiest to the hardest items reduces the likelihood of respondents leaving the test incomplete and increases the probability of correct responses (Anaya et al., 2022). The developed instrument adheres to this recommendation by starting with an easy item, followed by items of moderate difficulty. However, the difficult item appears as the eleventh question, followed by a moderate item as the twelfth question. This arrangement suggests the need to review and refine the test item sequence to align better with established recommendations.

The item difficulty levels are also displayed on the Wright map presented in Figure 3. The left side of the Wright map illustrates the difficulty of the test items. Item difficulty is expressed in logits, with a logit value of 0 arbitrarily set as the average level of item difficulty (Bond et al., 2020). Figure 3 shows a scale from -3.0 to 3.0, representing a range from very easy to very difficult. The lower the logit value, the easier the item is for most test takers, while the higher the logit value, the more difficult the item becomes. Therefore, items 6, 7, 9, 5, and 11 show positive logits that increase upwards, indicating increasingly difficult items. Item 12 has a logit of 0, representing the average difficulty level of the items. Items 2, 3, 10, 8, 4, and 1 display negative logits, scattered downward, making these items easier. Figure 3 indicates that all test items fall within an acceptable range of item difficulty indexes. Items with difficulty levels ranging from -2 to 2 are considered to have a good level of difficulty (Asriadi & Hadi, 2021; Sumaryanta, 2021).

The Wright map also provides insights into the distribution of respondents' abilities and the difficulty levels of the test items on the same scale. It helps evaluate the match between item difficulty and the respondents' abilities (Salman & Abd. Aziz, 2015). The right side of the map shows the distribution of student abilities, where the most difficult items (logit 3.0) were answered correctly by very few students, represented by three "X" marks, each symbolizing three students. Conversely, most students are located around difficulty level -1.0 and below (very easy). It indicates that as item difficulty decreases (negative logits), the number of correct responses increases, with most students performing well on easier items. This observation highlights the need to enhance physics instruction to improve students' conceptual understanding of mechanical waves. Strengthening their comprehension is critical to ensure a broader range of students can successfully tackle items with varying difficulty levels.

Implications and Limitations

The development of the instrument resulted in 12 test items designed to assess high school students' conceptual understanding of the topic of mechanical waves. The developed test was categorized as valid, reliable, and possessing an appropriate item difficulty index, indicating good item quality. This study's findings are significant for enriching the literature by providing a valid test instrument to measure high school students' conceptual understanding of physics in the context of mechanical waves, enabling further exploration of students' conceptions on this topic. Practically, the findings yield an instrument that teachers can utilize to assess or evaluate students' learning outcomes in physics. Additionally, the Rasch model analysis can serve as a guide for researchers and physics teachers in developing and evaluating the quality of test instruments.

The limitations of this study lie in the sample used for testing. The test sample was not selected based on respondents with low, medium, and high abilities. A diverse sample, encompassing a range of abilities, is crucial to ensure that the developed instrument is relevant and fair to all respondents. Furthermore, the test instrument was administered online, which minimized teacher supervision. This factor may have influenced the results of the instrument testing. Future research is recommended to ensure that test respondents represent various ability levels and that the test administration process is conducted under proper teacher supervision.

CONCLUSION

The development of the test instrument resulted in 12 items designed to investigate high school students' understanding of physics concepts on the topic of mechanical waves. Based on Rasch analysis, all 12 items met the infit and outfit criteria that align with the Rasch model, indicating that all items are acceptable for use in measurement. In terms of reliability, the test items exhibited high reliability, while the reliability of the individuals was moderate, influenced by the consistency of the respondents' answers. Additionally, all items had appropriate and acceptable difficulty levels. Overall, the developed test instrument demonstrates good quality and can be used to investigate high school students' understanding of mechanical wave concepts.

DISCLOSURE STATEMENT

The authors do not have any potential conflicts of interest to disclose.

FUNDING STATEMENT

This work does not receive funding.

ETHICS APPROVAL

There is no ethics approval needed because the research participants in this study were anonymized, and all data collected during the study were used only for the purpose of research. The results of the study would not cause any harm to the research participants.

REFERENCES

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2022). Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review*, *90*, 102293. https://doi.org/https://doi.org/10.1016/j.econedurev.2022.102293

Andrich, D., & Marais, I. (2019). Reliability and validity in classical test theory. In D. Andrich & I. Marais (Eds.), *A course in Rasch measurement theory: Measuring in the educational, social and health sciences* (pp. 41–53). Springer Nature Singapore. https://doi.org/10.1007/978-981-13-7496-8_4

Asriadi, M., & Hadi, S. (2021). Analysis of the quality of the formative test items for physics learning using the Rasch model in the 21st century learning. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, *6*(2), 158-166. https://doi.org/10.26737/jipf.v6i2.2030

Azizah, I., & Supahar, S. (2023). Analisis kualitas butir soal penilaian harian bersama I fisika kelas X SMA Negeri 1 Patikraja. *Jurnal Pendidikan Fisika*, *10*(2), 90–104. https://doi.org/10.21831/jpf.v10i2.18230

Azizah, N., Suseno, M., & Hayat, B. (2022). Item analysis of the Rasch model items in the final semester exam Indonesian language lesson. *World Journal of English Language*, *12*(1), 15–26. https://doi.org/10.5430/wjel.v12n1p15

Baker, F. B. (2001). *The basics of Item Response Theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Balta, N., Japashov, N., Salibašić Glamočić, D., & Mešić, V. (2022). Development of the high school wave optics test. *Journal of Turkish Science Education, 19*(1), 306-331. https://doi.org/10.36681/tused.2022.123

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch Model* (4th ed.). Routledge. https://doi.org/10.4324/9780429030499

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE Life Sciences Education*, *15*(4). https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. Springer Dordrecht. https://doi.org/10.1007/978-94-007-6857-4

Chen, L., Uemura, H., Hao, H., Goda, Y., Okubo, F., Taniguchi, Y., Oi, M., Konomi, S., Ogata, H., & Yamada, M. (2018). Relationships between collaborative problem solving, learning performance, and learning behavior in science education. *Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2018*, 17–24. https://doi.org/10.1109/TALE.2018.8615254

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, *24*(2), 97–140. https://doi.org/10.1080/10888691.2018.1537791

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education)*, *9*(1), 24–36. https://doi.org/10.21831/reid.v9i1.53514

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics - Physics Education Research*, *10*(2), 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Faradillah, A., & Febriani, L. (2021). Mathematical trauma students' junior high school based on grade and gender. *Infinity Journal*, *10*(1), 53–68. https://doi.org/10.22460/infinity.v10i1.p53-68

Goodhew, L. M., Robertson, A. D., Heron, P. R. L., & Scherr, R. E. (2019). Student conceptual resources for understanding mechanical wave propagation. *Physical Review Physics Education Research*, *15*(2), 020127. https://doi.org/10.1103/PhysRevPhysEducRes.15.020127

Halim, A., Suriana, S., & Mursal, M. (2017). Dampak problem based learning terhadap pemahaman konsep ditinjau dari gaya berpikir siswa pada mata pelajaran fisika. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, *3*(1), 1–10. https://doi.org/10.21009/1.03101

Habibi, H., Jumadi, J., & Mundilarto, M. (2019). The Rasch-rating scale model to identify learning difficulties of physics students based on self-regulation skills. *International Journal of Evaluation and Research in Education, 8*(4), 659–665. https://doi.org/10.11591/ijere.v8i4.20292

Hanna, W. F., & Retnawati, H. (2022). Analisis kualitas butir soal matematika menggunakan model Rasch dengan bantuan software Quest. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *11*(4), 3695. https://doi.org/10.24127/ajpm.v11i4.5908

Hope, D., Kluth, D., Homer, M., Dewar, A., Goddard-Fuller, R., Jaap, A., & Cameron, H. (2024). Exploring the use of Rasch modelling in "common content" items for multi-site and multi-year assessment. *Advances in Health Sciences Education, 30,* 427–438. https://doi.org/10.1007/s10459-024-10354-y

Istiyono, E. (2020). *Pengembangan instrumen penilaian dan analisis hasil belajar fisika dengan teori tes klasik dan modern*. UNY Press.

Kanyesigye, S. T., Uwamahoro, J., & Kemeza, I. (2022). Difficulties in understanding mechanical waves: Remediated by problem-based instruction. *Physical Review Physics Education Research*, *18*(1), 010140. https://doi.org/10.1103/PhysRevPhysEducRes.18.010140

Kola, A. J. (2017). Investigating the conceptual understanding of physics through an interactive lecture-engagement. *Cumhuriyet International Journal of Education*, *6*(1), 82–96. http://cije.cumhuriyet.edu.tr/en/pub/issue/29856/321440

Kurniawan, A., Istiyono, E., & Daeng Naba, S. (2024). Item quality analysis of physics concept understanding test with Rasch model. *JIPF (Jurnal Ilmu Pendidikan Fisika)*, *9*(3), 474–486. https://doi.org/10.26737/jipf.v9i3.5692

Kurniawan, F., Samsudin, A., Chandra, D. T., Sriwati, E., Zahran, M., Gani, A. W., Ramadhan, B. P., Aminudin, A. H., & Ramadani, F. (2023). Assessing conceptual understanding of high school students on transverse and stationary waves through Rasch analysis in Malang. *Journal of Physics: Conference Series*, *2596*(1). https://doi.org/10.1088/1742-6596/2596/1/012060

Kurpius, S. E. R., & Stafford, M. E. (2005). *Testing and measurement: A user-friendly guide*. Sage Publications.

Lafifa, F., & Dadan, R. (2024). Innovations in assessing students' digital literacy skills in learning science: Effective multiple choice closed-ended tests using Rasch model. *Turkish Online Journal of Distance Education*, *25*(3), 44–56. https://dergipark.org.tr/tr/download/article-file/3425765

Larasati, P. E., Supahar, & Yunanta, D. R. A. (2020). Validity and reliability estimation of assessment ability instrument for data literacy on high school physics material. *Journal of Physics: Conference Series*, *1440*(1), 012020. https://doi.org/10.1088/1742-6596/1440/1/012020

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes*. Mitra Cendekia.

Mešić, V., Neumann, K., Aviani, I., Hasović, E., Boone, W. J., Erceg, N., Grubelnik, V., Sušac, A., Glamočić, D. S., Karuza, M., Vidak, A., AlihodŽić, A., & Repnik, R. (2019). Measuring students' conceptual understanding of wave optics: A Rasch modeling approach. *Physical Review Physics Education Research*, *15*(1), 010115. https://doi.org/10.1103/PhysRevPhysEducRes.15.010115

Pals, F. F. B., Tolboom, J. L. J., & Suhre, C. J. M. (2023). Development of a formative assessment instrument to determine students' need for corrective actions in physics: Identifying students' functional level of understanding. *Thinking Skills and Creativity*, *50*, 101387. https://doi.org/10.1016/j.tsc.2023.101387

Planinic, M., Jelicic, K., Matejak Cvenic, K., Susac, A., & Ivanjek, L. (2024). Effect of an inquiry-based teaching sequence on secondary school students' understanding of wave optics. *Physical Review Physics Education Research*, *20*(1), 010156. https://doi.org/10.1103/PhysRevPhysEducRes.20.010156

Putranta, H., & Supahar. (2019). Development of physics-tier tests (PysTT) to measure students' conceptual understanding and creative thinking skills: A qualitative synthesis. *Journal for the Education of Gifted Young Scientists*, *7*(3), 747–775. https://doi.org/10.17478/jegys.587203

Putri, A. H., Sutrisno, S., & Chandra, D. T. (2020). Efektivitas pendekatan multirepresentasi dalam pembelajaran berbasis masalah untuk meningkatkan pemahaman konsep siswa SMA pada materi gaya dan gerak. *Journal of Natural Science and Integration*, *3*(2), 205–214. http://dx.doi.org/10.24014/jnsi.v3i2.9400

Rahim, A., Hadi, S., Susilowati, D., Marlina, & Muti'ah. (2023). Developing of Computerized Adaptive Test (CAT) based on a learning management system in mathematics final exam for junior high school. *International Journal of Educational Reform*. https://doi.org/10.1177/10567879231211297

Retnawati, H., & Wulandari, N. F. (2019). The development of students' mathematical literacy proficiency. *Problems of Education in the 21st Century*, *77*(4), 502–514. https://doi.org/10.33225/pec/19.77.502

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Kuram ve Uygulamada Egitim Bilimleri*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Salman, A., & Abd. Aziz, A. (2015). Evaluating user readiness towards digital society: A Rasch measurement model analysis. *Procedia Computer Science*, *65*, 1154–1159. https://doi.org/10.1016/j.procs.2015.09.028

Sartika, R. P. (2018). The implementation of problem-based learning to improve students' understanding in management of laboratorium subject. *EDUSAINS*, *10*(2), 197–205. https://doi.org/10.15408/es.v10i2.7376

Setyawarno, D. (2017). *Makalah PPM: Panduan Quest*. FMIPA UNY.

Shanti, M. R. S., Istiyono, E., Munadi, S., Permadi, C., Pattiserlihun, A., & Sudjipto, D. N. (2020). Analisa penilaian soal fisika menggunakan model Rasch dengan Program R. *Jurnal Sains dan Edukasi Sains*, *3*(2), 46–52. https://doi.org/10.24246/juses.v3i2p46-52

Sharma, V., Gupta, N. L., & Agarwal, A. K. (2023). Impact of ICT-enabled teaching–learning processes in physical sciences in Indian higher education in light of COVID-19: A comprehensive overview. *National Academy Science Letters*, *46*(5), 465–469. https://doi.org/10.1007/s40009-023-01225-y

Sumaryanta. (2021). *Teori Tes Klasik & Teori Respon Butir: Konsep & contoh penerapannya*. CV. Confident.

Szabó, G. (2008). *Applying Item Response Theory in language test item bank building*. Peter Lang. https://books.google.co.id/books?id=I0V9AAAAMAAJ

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*(2), 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Mesa Press.

Xie, L., Liu, Q., Lu, H., Wang, Q., Han, J., Feng, X. M., & Bao, L. (2021). Student knowledge integration in learning mechanical wave propagation. *Physical Review Physics Education Research*, *17*(2), 020122. https://doi.org/10.1103/PhysRevPhysEducRes.17.020122

Yim, L. W. K., Lye, C. Y., & Koh, P. W. (2024). A psychometric evaluation of an item bank for an English reading comprehension tool using Rasch analysis. *REID (Research and Evaluation in Education)*, *10*(1), 18–34. https://doi.org/10.21831/reid.v10i1.65284