

Evaluating the Quality of a Physics Problem-Solving Ability Instrument on Sound Wave Material Using the Rasch Measurement Model

Rahma Alliya Acquilla , Hera Novia* , and Nurdini Nurdini 

Physics Education Study Program, Universitas Pendidikan Indonesia, Jalan Dr. Setiabudi No.229, Bandung, West Java, Indonesia

* Corresponding author(s), e-mail: heranovia@gmail.com

Article Info:

Article History:

received: 26 June 2026

accepted: 17 September 2026

available online: 21 September 2026

Keyword:

physics problem-solving ability, rasch model, assessment instrument, sound waves, psychometric analysis

<https://doi.org/10.26877/lpt.v5i3.926>

Copyright © 2026 by author(s).

Published by Lontar Physics Today.

This is an open access article under the CC BY license (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



OPEN ACCESS

Abstract:

This study analyzed the psychometric characteristics of a physics problem-solving ability test on sound wave material using Many-Facet Rasch Measurement (MFRM) with the Partial Credit Model (PCM). In this descriptive quantitative study, an eight-item essay instrument based on Polya's four problem-solving stages (understanding the problem, devising a plan, carrying out the plan, and looking back) was administered to 74 Grade XI students at a public senior high school in Cimahi City (purposive sampling). Analysis covered unidimensionality, item and person reliability, item fit, item difficulty, discriminating power, category functioning, and differential item functioning (DIF). The raw variance explained by Rasch measures reached 65.45%, indicating very good unidimensionality. Seven items met the Rasch fit criteria, while Item 1a showed marginal fit and was retained. Item reliability was high (0.93) and person reliability was fair (0.71). Difficulty analysis classified one item as easy (1c), six as moderate (1a, 1b, 1d, 2a, 2c, 2d), and one as difficult (2b), indicating a similar target ability range across items. All items showed positive discrimination (Point Measure Correlation 0.37–0.71). Three items had disordered response-category thresholds requiring refinement, and DIF analysis identified four items functioning differently by gender, with DIF direction balanced between male and female students. These findings indicate satisfactory psychometric quality for measuring physics problem-solving ability on sound wave material, though several items warrant further qualitative review. Further development is recommended, including expanding item numbers, refining response categories, and testing with larger, more diverse samples to improve generalizability.

1. Introduction

Problem-solving plays an important role in physics learning because it helps students construct new knowledge and facilitates the learning process (Mukhopadhyay, 2013). In physics, problem-solving ability is not merely a general skill but involves a series of cognitive processes. One widely used framework for describing this process is Polya (1957) four-stage model, which consists of understanding the problem, devising a plan, carrying out the plan, and looking back. Physics learning requires students not only to understand abstract concepts but also to apply them in real and complex situations (Putra et al., 2025). Similarly, Musengimana et al. (2025) emphasized that problem-solving ability can strengthen conceptual understanding while developing students' capacity to address real-world problems. In this context, physics problem solving requires students to coordinate conceptual understanding, mathematical reasoning, and different forms of representation throughout the problem-solving process.

These demands are particularly evident in the topic of sound waves. Sound wave concepts are inherently abstract because the phenomena are dynamic and cannot be directly observed, making them difficult for students to understand through verbal explanations or symbolic representations alone (Frank et al., 2025). To solve problems related to sound waves, students must construct appropriate mental representations and connect them with relevant mathematical relationships. However, students often experience difficulties when operating wave equations and connecting mathematical representations with verbal and graphical representations (Amiroh et al., 2021). These challenges can occur throughout the

problem-solving process. Students must first identify relevant physical quantities from the given information, determine an appropriate relationship among those quantities, apply the selected procedure, and finally evaluate whether the obtained result is consistent with the physical situation. Therefore, the abstract and representational characteristics of sound waves make this topic a particularly demanding context for examining students' problem-solving abilities.

Despite the importance of problem-solving ability, previous studies have shown that many high school students still demonstrate relatively low performance in physics problem solving (Listiawan et al., 2026; Nurrohmah et al., 2026). For example, Putra et al. (2025) found that most students were in the moderate to low categories, particularly in the stages of devising a plan, carrying out the plan, and looking back. These findings suggest that students' difficulties may not occur uniformly throughout the problem-solving process. Consequently, identifying students' abilities at each stage is important for understanding where their difficulties are most pronounced and for providing a basis for improving physics instruction. However, such identification requires an instrument that can assess students' performance across different stages of problem solving, particularly in the context of sound waves.

The identification of students' problem-solving abilities therefore depends on the availability of valid and reliable measurement instruments. In addition to measuring students' ability profiles, assessment instruments can also support the development of higher-order thinking skills, including problem-solving ability (Lestari et al., 2019). Information obtained from high-quality instruments can provide teachers with a more accurate understanding of students' abilities and support more appropriate instructional decisions. Therefore, developing and evaluating the quality of problem-solving instruments is an important step in efforts to improve physics learning.

Several Rasch-based problem-solving instruments have been developed in physics education. However, these instruments have generally been designed for specific content areas, problem-solving frameworks, and participant groups. At the junior high school level, (Rokhmat & Hanif, 2017) developed an instrument on vibration, waves, and sound based on (J. Docktor et al., 2007) five indicators. This framework includes important aspects of physics problem solving, such as visualizing the problem, applying physics principles, and performing mathematical procedures. However, it does not explicitly include a stage for evaluating the reasonableness of the final solution.

Similar approaches have also been applied in other contexts. At the senior high school level, Listiawan et al. (2026) applied a Heller-based framework to the topic of heat. Nurrohmah et al. (2026), meanwhile, developed a Rasch-based instrument on alternative energy using Treffinger & Isaksen (2004) Creative Problem Solving framework. This framework places greater emphasis on creative and divergent thinking than on the systematic integration of conceptual and mathematical reasoning typically required in physics problem solving. Ridwan et al. (2023) also employed Rasch analysis in developing a problem-solving instrument on climate change. However, its indicators—defining the problem, identifying root causes, generating alternative solutions, and selecting the best solution—represent a more general problem-solving process for socio-scientific issues. Moreover, the study involved university students rather than high school students.

These previous studies demonstrate the potential of Rasch analysis for evaluating problem-solving instruments while also highlighting differences in the frameworks, content areas, and target populations used. However, no previous Rasch-based study has adopted Polya's (1957) four-stage framework, consisting of understanding the problem, devising a plan, carrying out the plan, and looking back. The inclusion of the looking-back stage is particularly important because it requires students to reflect on and

evaluate their solutions, thereby capturing an aspect of the problem-solving process that is closely related to metacognitive evaluation. In addition, previous studies have not applied Polya's framework to sound waves, a topic that requires students to coordinate conceptual understanding with mathematical reasoning and multiple representations. This creates a gap in the development of a physics problem-solving instrument that combines Polya's complete problem-solving stages with Rasch analysis for high school students in the context of sound waves.

To address this gap, the present study develops and evaluates a Rasch-based physics problem-solving instrument based on Polya's framework for the topic of sound waves. In contrast to existing instruments that emphasize particular aspects of physics problem solving, creative and divergent thinking, or more general problem-solving processes, the proposed instrument covers the process from understanding a problem to evaluating the resulting solution. This structure allows students' performance to be examined across different stages of problem solving while maintaining a clear connection to the characteristics of physics content.

The instrument focuses on two sound wave subtopics, namely string frequency and sound intensity level. These subtopics involve different mathematical relationships and require students to work with abstract representations as well as quantitative procedures. Through Rasch analysis, the instrument can be evaluated not only in terms of overall item difficulty but also in terms of how item difficulty is distributed across Polya's problem-solving stages and the two content areas. Such information can help identify the aspects of problem solving that are most challenging for high school students when solving problems related to sound waves.

Based on the above discussion, this study aims to analyze the quality of a physics problem-solving ability test instrument on sound waves for high school students using the Rasch model. The results are expected to provide information about the psychometric characteristics of the instrument and contribute to the development of higher-quality instruments for assessing students' physics problem-solving abilities.

2. Theoretical Framework

2.1. *Physics Problem-Solving Ability*

Problem-solving ability is a complex competency that enables individuals to address challenges by drawing on their prior knowledge and engaging in higher-order cognitive processes, such as analysis, evaluation, and reflection Graesser et al. (2018). Although definitions of problem solving generally involve identifying a problem and finding an effective solution, scholars differ in how they describe the process. Özreçberoğlu & Çağanağa (2018), for example, primarily describe problem solving as a decision-making process, whereas Dörner & Funke, (2017) conceptualize complex problem solving as a set of self-regulated psychological processes and activities through which individuals overcome barriers between a given state and a desired goal state, particularly in complex, dynamic, and ill-defined situations. These different perspectives have implications for the way problem-solving ability is assessed. A decision-making perspective tends to emphasize the selection of available alternatives, whereas a staged approach focuses more on the process of constructing a solution. The latter is particularly relevant to physics, where students are often required to develop a solution through conceptual and quantitative reasoning rather than simply select from predetermined alternatives. For this reason, frameworks that describe problem solving as a sequence of interconnected stages, such as Polya's framework discussed in Section 2.2, are particularly relevant to physics education research.

Problem-solving ability is especially important in physics education, where it is widely recognized as a central learning goal and a major area of research in physics education research (PER) (J. L. Docktor & Mestre, 2014). Solving physics problems requires more than following memorized procedures. Students must analyze the problem situation, identify relevant physical principles, and coordinate conceptual understanding with mathematical reasoning to solve complex and non-routine problems (Maloney, 2011). Given these demands, assessing students' problem-solving ability is important not only for evaluating learning outcomes but also for identifying where students experience difficulties during the problem-solving process. Such assessment can provide information about whether students' difficulties are primarily related to conceptual understanding, procedural execution, or the evaluation of their solutions. This diagnostic function underlies the development of the problem-solving instrument examined in the present study.

2.2. Polya's Problem-Solving Framework in Physics Education

Polya's (1957) four-stage framework—understanding the problem, devising a plan, carrying out the plan, and looking back—was originally developed in the context of mathematical problem solving. Therefore, its application to physics problem solving requires careful consideration of its relevance to the characteristics of physics learning. The connection between Polya's framework and physics problem solving, however, has a well-established history. Maloney's (2011) overview of problem-solving research in physics education traces this connection through the Minnesota structured problem-solving strategy developed by Heller & Holla Baugh (1992), which was explicitly based on Polya's four-step framework. In adapting the framework to physics, the strategy was developed into five physics-oriented steps: focusing the problem, describing the physics, planning the solution, executing the solution, and evaluating the answer. This adaptation shows that Polya's general problem-solving framework can be meaningfully applied to physics by accommodating the conceptual and mathematical demands that characterize physics problems.

Although Polya's framework has influenced the development of problem-solving approaches in physics, it differs in emphasis from frameworks that were specifically designed for physics education. For example, Docktor and Heller's (2009) five-indicator rubric—useful description, physics approach, application of physics, mathematical procedure, and logical progression—provides a more detailed description of the processes involved in solving physics problems. Its indicators are particularly useful for examining how students identify relevant information, apply physical principles, and carry out mathematical procedures. However, the framework does not separate the evaluation of the final solution as a distinct stage of the problem-solving process. In contrast, Polya's looking-back stage explicitly requires students to review and evaluate the solution they have obtained, including whether it is reasonable in relation to the original problem.

This reflective component is particularly relevant to the nature of expert problem solving. Research in physics education has emphasized the importance of metacognitive regulation, including planning, monitoring, and evaluating one's own thinking, in distinguishing expert problem solvers from novices (Maloney, 2011; Schoenfeld, 1985, as cited in Maloney, 2011). From this perspective, Polya's framework provides an advantage for the present study because it includes evaluation as an explicit part of the problem-solving process. Rather than focusing only on how students construct and execute a solution, the framework also makes it possible to examine whether students reflect on the appropriateness of the solution they obtain. This consideration forms the basis for selecting Polya's framework as the conceptual foundation for the instrument developed in this study.

Each of Polya's four stages also involves a distinct cognitive demand. Understanding the problem requires students to interpret the given information and identify relevant physical quantities from verbal, graphical, or other representations. Devising a plan involves determining an appropriate strategy and selecting the physical or mathematical relationships needed to solve the problem. Carrying out the plan requires students to implement the selected procedure while maintaining consistency with the underlying physical concepts. Finally, looking back requires students to evaluate whether the resulting answer is reasonable and physically meaningful. Together, these stages provide a structured basis for developing an instrument that examines students' performance throughout the problem-solving process rather than treating problem-solving ability as a single, undifferentiated skill.

2.3. Sound Waves as a Content for Problem-Solving Assessment

The topic of sound waves provides a particularly demanding context for examining students' problem-solving abilities. Sound wave phenomena are inherently abstract because they are dynamic and cannot be directly observed, making them difficult for students to understand through verbal explanations or symbolic representations alone (Frank et al., 2025). To solve problems involving sound waves, students must construct appropriate mental representations and connect these representations with relevant mathematical relationships. However, students often experience difficulties when operating wave equations and relating mathematical representations to verbal and graphical representations (Amiroh et al., 2021). These characteristics make sound waves a relevant context for examining how students move through the different stages of problem solving.

Sound Wave Content. In this study, the sound wave topic is represented through two sub-concepts, namely string frequency (*frekuensi dawai*) and sound intensity level (*taraf intensitas bunyi*). Although both sub-concepts are included within the broader topic of sound waves, they involve different mathematical relationships and therefore require different forms of conceptual and procedural reasoning.

The first sub-concept, string frequency, involves standing waves formed when a string is plucked. These waves result from the superposition of incident and reflected waves between two fixed ends. The fundamental frequency and higher harmonics differ in their wavelengths, while the frequency of the n -th harmonic depends on the wave speed and the length of the string. In turn, wave speed is determined by the tension and linear mass density of the string. This relationship requires students to connect several physical quantities rather than directly apply a single formula. Students must recognize how wave speed relates to tension and mass density before connecting it to the frequency of the string. Such reasoning places particular demands on the stages of devising a plan and carrying out the plan.

The second sub-concept, sound intensity level, involves the relationship between sound intensity and the threshold of human hearing. In addition, changes in sound intensity with distance from a source follow an inverse-square relationship. Consequently, students must work with both logarithmic and inverse-square relationships when solving problems related to sound intensity level. This differs from the reasoning required for string frequency, which primarily involves relationships among physical quantities that can be expressed through direct or inverse proportionality. Problems involving sound intensity level therefore require students to adapt their solution strategies to a different mathematical structure rather than applying the same reasoning used in other sound wave problems.

Together, these two sub-concepts allow the instrument to examine students' problem-solving performance across different types of quantitative reasoning within the same general topic. String frequency requires students to coordinate relationships among several physical quantities, whereas sound intensity level requires them to interpret logarithmic and inverse-square relationships. The inclusion of

both sub-concepts therefore makes it possible to examine whether the difficulty of solving problems differs not only across the stages of Polya's framework but also across different mathematical structures within the topic of sound waves.

Problem-Solving Indicators. To examine students' performance across these stages, each item in the instrument is assessed using a scoring rubric that operationalizes Polya's four stages into observable indicators, following the partial-credit structure described in Section 2.4. Understanding the problem is indicated by students' ability to correctly identify the known and unknown quantities presented in verbal or graphical form and is scored on a three-point scale. Devising a plan refers to the ability to select an appropriate physical relationship or formula connecting the known and unknown quantities and is scored on a two-point scale. For example, students may need to determine the appropriate relationship for calculating wave speed before using it to determine the frequency of a string, or identify the relevant relationship between sound intensity and distance. Carrying out the plan refers to the correctness and consistency of the mathematical procedure, including appropriate substitution and unit handling, and is scored on a three-point scale. Finally, looking back refers to students' ability to evaluate whether the obtained result is physically reasonable and, where relevant, express the final answer using appropriate units; this stage is scored on a two-point scale.

The same four-indicator structure is applied to both sub-concepts. This consistency makes it possible to compare students' performance at each stage of problem solving while also examining whether the difficulty associated with each stage differs according to the type of quantitative reasoning required by the problem.

2.4. Rasch Measurement Model

The Rasch model is a psychometric measurement model that estimates person ability and item difficulty on a common interval scale. By transforming ordinal response data into logit measures, the model provides a more objective basis for interpreting assessment results (Irribarra, 2018). Traditionally, many assessment instruments have been evaluated using Classical Test Theory (CTT). Although CTT provides useful information about test reliability and item performance, its estimates are generally dependent on the particular sample and set of items used. Consequently, item parameters and respondent scores may vary across different populations. In addition, CTT primarily relies on total scores, which provides limited information about the relationship between individual item difficulty and respondent ability (Ramadhani & Handayani, 2025; Sumintono & Widhiarso, 2015). These limitations make the Rasch model particularly relevant for evaluating an instrument designed to examine students' performance across different stages of problem solving.

In the present study, the Rasch model is applied using the Partial Credit Model (Masters, 1982). The selection of the PCM is based on the polytomous scoring structure of the instrument. Each item is scored using indicators derived from Polya's problem-solving stages, with different maximum scores assigned to the indicators. Understanding the problem and carrying out the plan are scored up to three points, whereas devising a plan and looking back are scored up to two points. This scoring structure produces items with different numbers of response categories. The Rating Scale Model (Andrich, 1978), by contrast, assumes a common rating-scale structure across items, including shared step or threshold parameters. Such an assumption is more appropriate when all items have the same response-category structure. Because the indicators in this instrument have different maximum scores and, consequently, different category structures, the PCM is more appropriate because it allows the step difficulties to vary across items.

The use of the Rasch model is also consistent with the diagnostic purpose of the instrument. This study is intended not only to assess students' overall problem-solving ability but also to identify the relative difficulty of problems associated with different stages of the problem-solving process. The Rasch model places person ability and item difficulty on the same logit scale, allowing their relationship to be examined directly. This relationship can be visualized through a Wright map, which provides information about the distribution of students' abilities and item difficulties (Mulyani et al., 2017). As a result, the analysis can provide a more detailed picture of which items, and consequently which aspects of problem solving, are relatively more challenging for students.

Within the Rasch framework, item discrimination is examined using the Point Measure Correlation (Pt Measure Corr.), which indicates the relationship between item performance and the overall latent construct being measured. Its interpretation, however, differs from the use of item-discrimination statistics in Classical Test Theory. In CTT, point-biserial correlations are often evaluated using minimum thresholds, such as 0.30 or 0.40. In Rasch analysis, a positive Point Measure Correlation is primarily expected because it indicates that higher performance on an item is generally associated with higher levels of the measured ability. However, item quality should not be determined solely by the magnitude of this correlation. Other Rasch indicators, particularly mean-square fit statistics, provide important information about the extent to which an item functions consistently with the expectations of the measurement model (Boone & Staver, 2020; Linacre, 2002). Therefore, Point Measure Correlation values in this study are interpreted within the Rasch measurement framework rather than according to discrimination thresholds derived from CTT.

2.5. Conceptual Framework

Taken together, these theoretical foundations provide a coherent conceptual basis for the present study. Physics problem-solving ability, as discussed in Section 2.1, is examined through Polya's four-stage framework (Section 2.2): understanding the problem, devising a plan, carrying out the plan, and looking back. This framework was selected because it includes an explicit stage for evaluating the solution, which provides an important reflective dimension within the problem-solving process.

The framework is applied to the topic of sound waves (Section 2.3), which provides a challenging context for examining students' problem-solving abilities. The two sub-concepts included in the instrument, namely string frequency and sound intensity level, involve different physical and mathematical relationships. This structure makes it possible to examine problem difficulty not only across the four stages of problem solving but also across different types of quantitative reasoning within the same topic.

Finally, the resulting instrument is evaluated using the Rasch measurement model (Section 2.4). The use of a polytomous Rasch model is consistent with the partial-credit scoring structure of the instrument and allows person ability and item difficulty to be examined on a common measurement scale. Together, these theoretical and methodological foundations provide a basis for obtaining detailed information about students' performance and identifying the aspects of physics problem solving that present the greatest difficulty when students solve problems related to sound waves.

3. Method

3.1. Research Design

This study employed a descriptive quantitative approach to examine the quality of a measurement instrument designed to assess students' problem-solving ability on the topic of sound waves. This

approach was selected because the study focused on systematically evaluating the quality of the instrument based on empirical data obtained through direct testing with students.

The research was conducted in five stages, as illustrated in Figure 1. The first stage was a preliminary study aimed at obtaining an initial picture of students' problem-solving abilities and identifying issues relevant to the development of the instrument. A preliminary problem-solving test was administered to students at a public high school in Cimahi City during a 45-minute class period. The second stage involved developing the instrument based on Polya's four problem-solving indicators and contextualizing the items within the topic of sound waves. This process was conducted over two weeks and involved drafting items according to the four indicators and the relevant sound wave content. The initial draft was subsequently reviewed internally by the research team before being submitted for expert validation. The third stage consisted of expert validation. The draft instrument was evaluated by four validators, comprising two physics education lecturers and two physics teachers. Each validator independently assessed the content, construct, and language appropriateness of the instrument. The feedback obtained from the validators was compiled and used as the basis for revising the items before the pilot test. In the fourth stage, the revised instrument was pilot-tested with 74 eleventh-grade students. The test was administered under supervised conditions, with students given 45 minutes to complete the instrument. Their responses were subsequently coded according to the predetermined polytomous scoring rubric. Finally, the coded response data were analyzed using the Rasch Partial Credit Model to evaluate the psychometric quality and functioning of the instrument.

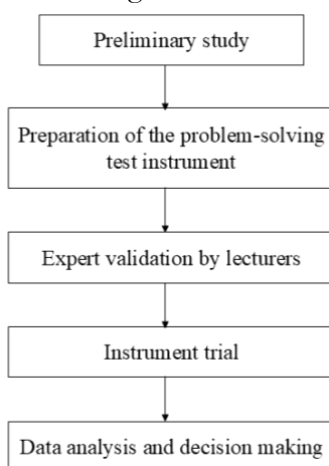


Figure 1. *Research stages*

3.2. Preliminary Study

Prior to developing the main instrument, a preliminary study was conducted to obtain an initial picture of students' problem-solving ability on the topic of sound waves. A preliminary problem-solving test developed by the researchers was administered to 68 eleventh-grade students from two classes at the same public high school in Cimahi City, selected through purposive sampling. The students obtained an average problem-solving score of 55.58 out of 100, indicating that their problem-solving ability on sound waves was still relatively low. This finding provided an empirical basis for identifying the research problem and highlighted the need for a valid and reliable instrument that could provide a more detailed picture of students' problem-solving ability on this topic.

3.3. Research Subject

The research participants were 74 eleventh-grade students from a public high school in Cimahi City. The participants were selected using purposive sampling based on their suitability for the study, as the sound wave material was taught at their grade level. The study was conducted in a single school because the school provided access to the target student population and allowed data collection to be carried out consistently across the preliminary and pilot phases. Using the same school also helped reduce potential variation associated with differences in instructional contexts across schools. Prior to data collection, formal permission was obtained from Universitas Pendidikan Indonesia and the school principal. All participating students provided informed consent before taking part in the study.

3.4. Instrument

The instrument used in this study was an essay-based problem-solving ability test consisting of two main problems. Each problem was designed to assess all four of Polya's (1957) problem-solving indicators: (1) understanding the problem, (2) devising a plan, (3) carrying out the plan, and (4) looking back. Thus, each main problem consisted of four sub-items, resulting in eight scored sub-items in total (see **Table 1**, Items 1a–1d and 2a–2d). The instrument was developed by the researchers based on Polya's four problem-solving indicators and contextualized within the topic of sound waves and the students' learning context. Each of the four Polya indicators was represented across the two main problems, allowing students' performance to be examined consistently across the problem-solving process. Each sub-item was scored using a polytomous scoring rubric developed to reflect different levels of students' performance at each problem-solving stage. The scoring rubric used for coding students' responses is provided in Appendix.

3.5. Expert Validation

Prior to pilot testing, the instrument was reviewed by four validators to assess the appropriateness of its content, construct, and language. Among the two physics education lecturers, the first specializes in Rasch-based instrument measurement, while the second combines expertise in instrument measurement with content expertise in sound wave material. Among the two physics teachers, the first is an active classroom teacher who has directly taught sound wave material to senior high school students in a formal school setting, while the second is an active physics teacher whose sound-wave teaching experience comes primarily from olympiad coaching and private tutoring rather than formal classroom instruction.

Table 1. *Aiken's V Validity Coefficients*

No.	V	Interpretation
1a	0.95	valid
1b	0.95	valid
1c	0.95	valid
1d	0.95	valid
2a	0.96	valid
2b	0.96	valid
2c	0.98	valid
2d	0.96	valid

The validity of each sub-item was evaluated using Aiken's V coefficient. Based on the Aiken's V critical value for four validators and a four-point rating scale, a coefficient of 0.94 was used as the minimum criterion for validity (Aiken, 1985). Each Aiken's V coefficient in Table 1 represents the averaged score across three validated aspects: language clarity, construct alignment, and content appropriateness. All eight sub-items obtained Aiken's V values ranging from 0.95 to 0.98, exceeding the specified criterion (see Table 1). The validators' feedback mainly concerned minor revisions to wording and language clarity.

These suggestions were incorporated into the final version of the instrument before it was administered in the pilot test.

3.6. Data Collection

Data were collected through a written essay test designed to assess students' physics problem-solving ability on the topic of sound waves. The test was administered face-to-face in the classroom during a physics lesson scheduled in coordination with the school. The test was conducted under consistent classroom conditions to ensure that all students completed the assessment under comparable circumstances.

Before the test began, the researcher verbally explained the instructions, including how to respond to the open-ended questions and the time allocated for the test. Students were instructed to show their complete problem-solving process in a logical sequence, beginning with identifying the given and unknown information, followed by devising a solution strategy, carrying out the necessary calculations, and finally reviewing the solution and its results. These instructions were provided to ensure that students' responses reflected the four stages of Polya's problem-solving framework and could be scored using a structured, process-based rubric (Docktor et al., 2016).

3.7. Data Analysis

The data obtained from the pilot study were analyzed using the Rasch Partial Credit Model (PCM) because the instrument used polytomous scoring for open-ended problem-solving responses. Rasch analysis was selected because it allows item difficulty and student ability to be estimated on a common logit scale and provides diagnostic information about item and person functioning (Sumintono & Widhiarso, 2015; Susdelina et al., 2018). The analysis was conducted using Minifacet software version 3.83.6 and included unidimensionality, person and item reliability and separation, item fit, item discrimination, item difficulty, Wright Map targeting, category functioning, differential item functioning (DIF), and person fit.

Unidimensionality. First, the dimensionality of the instrument was examined to determine whether the items primarily measured a single underlying construct, namely students' physics problem-solving ability. The analysis considered the percentage of raw variance explained by Rasch measures and the unexplained variance in the first contrast. The criteria used to evaluate unidimensionality are presented in **Table 2**.

Table 2. *Unidimensionality Criteria (Sumintono & Widhiarso, 2015)*

Variance explained by Rasch measures	Interpretation
$20\% \leq X \leq 40\%$	Sufficient
$40\% < X \leq 60\%$	Good
$X > 60\%$	Very Good

Person and Item Reliability. Second, person and item reliability were examined to evaluate the consistency of the estimated person measures and item difficulty hierarchy, respectively. Person reliability indicates the consistency of the ordering of students according to their estimated ability, whereas item reliability indicates the consistency of the estimated item difficulty hierarchy. Person and item separation were also examined to determine the extent to which the instrument could distinguish different levels of student ability and item difficulty. The interpretation criteria for reliability are presented in Table 3.

Table 3. *Criteria for Person Reliability and Item Reliability (Sumintono & Widhiarso, 2014)*

Person reliability and item reliability	Interpretation
$X > 0.94$	Excellent
$0.91 \leq X < 0.94$	Very Good
$0.81 \leq X < 0.91$	Good

$0.67 \leq X < 0.81$	Fair
$X < 0.67$	Weak

Item Fit. Third, item fit was examined to determine whether each test item functioned as expected by the Rasch model. Item fit was evaluated using Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt Measure Corr.). The criteria used to evaluate item fit are presented in **Table 4**. An item was considered to function adequately when at least two of the three criteria were met (see **Table 5**), following Fisher (2007).

Table 4. *Acceptance Criteria for Outfit MNSQ, Outfit ZSTD, and Pt Measure Correlation (Sumintono & Widhiarso, 2015)*

Value	Criteria
Outfit MnSq	$0.5 < \text{MnSq} < 1.5$
Outfit ZStd	$-2.0 < \text{ZStd} < +2.0$
PT Measure Correlation	$0.4 < \text{PT Measure Correlation} < 0.85$

Table 5. *Interpretation of Item Quality (Fisher, 2007)*

Number of Criteria Met	Criteria
2 or 3	Used
0 or 1	Not Used

Discrimination. Fourth, item discrimination was evaluated using Point Measure Correlation (Pt Measure Corr.), which indicates the extent to which responses to an item are consistent with the overall ordering of students on the measured construct (Boone & Staver, 2020; Linacre, 2008). Positive values indicate that higher-performing students tend to obtain higher scores on the item. Pt Measure Corr. was interpreted together with the item-fit statistics rather than used as the sole criterion for determining item quality.

Item Difficulty Level. Fifth, item difficulty was examined using the logit measure values reported in the Item Measure table. The difficulty of each item was interpreted by comparing its logit measure with the mean and standard deviation of the item measures, following Sumintono and Widhiarso (2015). The classification criteria are presented in

Table 6.

Table 6. *Item difficulty level criteria (Sumintono & Widhiarso, 2015)*

Difficulty Level Criteria	Criterion
$\text{ME} < -1 \text{ SD}$	Easy
$-1 \text{ SD} \leq \text{ME} \leq +1 \text{ SD}$	Moderate
$\text{ME} > +1 \text{ SD}$	Difficult

Wright Map. The distribution of item difficulty and student ability was further examined using the Wright Map, which places both estimates on the same logit scale (Sumintono & Widhiarso, 2015). The Wright Map was used to evaluate the targeting of the instrument and identify portions of the ability continuum that were insufficiently represented by the test items (Mulyani et al., 2017).

Category Functioning. The functioning of the response categories within each polytomous item was examined using Rasch-Andrich thresholds. Thresholds represent the points on the logit scale at which adjacent response categories become equally probable. The categories were considered to function appropriately when the thresholds increased monotonically across successive categories. Disordered thresholds indicate that adjacent response categories may not represent distinct levels of performance along the ability continuum (Linacre, 2002; Bond & Fox, 2007).

Differential Item Functioning (DIF). DIF analysis was conducted to determine whether item difficulty differed systematically between male and female students after accounting for differences in

overall ability. DIF was examined by comparing the item difficulty estimates obtained separately for the two gender groups. Following Draba (1977), an item was identified as exhibiting significant DIF when the absolute contrast between the group-specific item difficulty estimates exceeded 0.5 logits and the absolute t-statistic was ≥ 1.96 ($p < .05$). Both criteria were required for an item to be classified as exhibiting significant DIF.

Person Misfit. Finally, person-fit analysis was conducted to examine the extent to which individual students' response patterns were consistent with the expectations of the Rasch model. The analysis used Infit and Outfit MNSQ statistics and their standardized values (ZSTD). Infit MNSQ is more sensitive to response patterns on items close to a student's estimated ability, whereas Outfit MNSQ is more sensitive to unexpected responses given the student's ability and item difficulty. MNSQ values between 0.5 and 1.5 were considered productive for measurement, whereas values exceeding 1.5 were considered indicative of potential misfit and were examined further.

4. Result

4.1. Unidimensionality

The dimensionality of the instrument was examined using Minifac to determine whether the items primarily measured a single underlying construct, namely students' physics problem-solving ability. The analysis showed that the raw variance explained by Rasch measures was 65.45%. The 65.45% raw variance explained by Rasch measures may be related to the shared conceptual and mathematical demands across the two sound-wave sub-topics and the four Polya problem-solving stages. Although string frequency and sound intensity level involve different mathematical relationships, both require students to interpret physical information, select appropriate relationships, carry out quantitative procedures, and evaluate the resulting solution. This result indicates that the instrument predominantly measured the intended underlying construct. The criteria used to interpret the raw variance explained by measures are presented in Table 2. The unidimensionality results obtained from the Item Dimensionality output in Many-Facets Rasch Model are presented in Table 7.

Table 7. *Item Unidimensionality*

Measure	Percentage
Variance explained by Rasch measures	65.45%
Variance of residuals	34.55%

4.2. Item and Person Reliability

Item Reliability. The reliability analysis showed an item reliability value ranging from 0.91 to 0.93, which indicates a high level of consistency in the ordering of item difficulty estimates. This result suggests that the hierarchy of item difficulties was estimated consistently and would be expected to remain relatively stable across comparable samples. The item reliability results obtained from the Many-Facets Rasch Model output are presented in Table 8.

Table 8. *Item Separation, Strata, and Reliability*

Statistic	Population	Sample
Separation	3.28	3.53
Strata	4.71	5.03
Reliability	0.91	0.93

Person Reliability. The Rasch analysis also yielded a person reliability value of 0.71. This indicates a fair level of consistency in the ordering of students according to their estimated problem-solving ability. The analysis also produced a person separation index of 1.56 and a person strata value of 2.41, indicating that the instrument was able to distinguish students into approximately two levels of ability. Extreme scores,

including perfect and zero scores, were excluded from the person reliability calculation because such scores do not provide sufficient information for precise estimation of person measures. The person reliability results obtained from the Many-Facets Rasch Model output are presented in Table 9.

Table 9. *Person Separation, Strata, and Reliability*

Statistic	Population	Sample
Separation	1.56	1.57
Strata	2.41	2.43
Reliability	0.71	0.71

4.3. Item Fit, Item Discrimination, and Item Difficulty Level

Item Fit. The empirical functioning of each test item was examined using Rasch fit statistics, namely Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt Measure Corr.). The acceptance criteria were based on Fisher (2007), with Outfit MNSQ values between 0.5 and 1.5, Outfit ZSTD values between -2.0 and $+2.0$, and positive Pt Measure Corr. values within the specified range presented in Table 4. The item difficulty measures, standard errors, and fit statistics are presented in

Table 10.

Table 10. *Interpretation of Item Validity for Problem-Solving Ability Test*

Test item	Measure (Logit)	SE	Outfit		PT-Measure Correlation	Fit Decision
			MnSq	ZStd		
1a	-0.36	0.22	1.35	2.0	0.37	Marginal fit; retained
1b	0.05	0.27	1.30	1.3	0.59	Fit
1c	-1.15	0.23	1.08	0.3	0.46	Fit
1d	-0.67	0.21	1.26	1.2	0.66	Fit
2a	-0.16	0.14	0.71	-0.7	0.71	Fit
2b	1.03	0.22	0.73	-1.8	0.62	Fit
2c	0.35	0.14	0.65	-1.0	0.69	Fit
2d	0.91	0.18	0.62	-1.4	0.61	Fit

The standard errors of the item difficulty estimates ranged from 0.14 to 0.27 logits. All eight items met the Outfit MNSQ criterion, with values ranging from 0.62 to 1.35. Seven items met the Outfit ZSTD criterion, with values ranging from -1.8 to 1.3 . Item 1a showed an Outfit MNSQ of 1.35 and a ZSTD value of 2.0, indicating a marginal fit with respect to the specified fit criteria. However, its Pt Measure Corr remained positive (0.37), indicating a consistent relationship between responses to the item and the overall measure. Considering these results together with the substantive relevance of the item to the targeted problem-solving construct, Item 1a was retained for further analysis.

The remaining seven items met the specified fit criteria. Overall, the fit statistics indicated that seven of the eight items functioned adequately according to the specified Rasch criteria, while Item 1a showed marginal fit and warranted further consideration. The standard errors were relatively comparable across items, although Item 1b had the highest SE at 0.27 logits. This value did not coincide with an item-fit violation. Thus, the item difficulty estimates were considered sufficiently precise for the subsequent analysis.

Item Discrimination. Item discrimination was further examined using the Pt Measure Corr. values reported in Table 10. In Rasch measurement, a positive Pt Measure Corr. indicates that students with higher overall ability tend to obtain higher scores on the corresponding item, suggesting that the item functions in the expected direction of the measured construct (Boone & Staver, 2020; Linacre, 2008).

All eight items produced positive Pt Measure Corr. values, ranging from 0.37 for item 1a to 0.71 for item 2a. This consistent positive direction indicates that the scoring patterns of all items were aligned

with the overall problem-solving ability measured by the instrument. Although item 1a produced the lowest value, its positive correlation and acceptable Outfit MNSQ and ZSTD values supported its retention. Nevertheless, the relatively lower Pt Measure Corr. of item 1a indicates that this item warrants attention in the interpretation of its functioning.

Item Difficulty Level. Item difficulty represents the level of ability required for students to obtain higher scores on an item and is expressed as a logit measure on the same scale as person ability (Sumintono & Widhiarso, 2015). Based on the item measure values presented in Table 10, item difficulty ranged from -1.15 logits for item 1c to 1.03 logits for item 2b, producing a total range of 2.18 logits across the eight items.

Following the classification criteria proposed by Sumintono and Widhiarso (2015), items with measure values below -1 SD were classified as easy, items between -1 SD and $+1$ SD were classified as moderate, and items above $+1$ SD were classified as difficult. Based on these criteria, item 1c was classified as easy, six items (1a, 1b, 1d, 2a, 2c, and 2d) were classified as moderate, and item 2b was classified as difficult. Thus, the instrument covered a range of difficulty levels, although most items were concentrated in the moderate category.

4.4. Wright Map

The Wright Map generated from the Rasch analysis and presented in Table 11 displays student ability and item difficulty on the same logit scale. The item locations provide a common reference for interpreting the distribution of student ability, while the detailed analysis of item difficulty is presented in the corresponding item-difficulty section. In terms of student ability, Students 12, 3, 37, and 5 were positioned at approximately $+3$ logits, representing the highest ability estimates in the sample. In contrast, Students 46, 50, 51, 52, 58, and 64 were located at approximately -2 logits, representing the lowest ability estimates. The remaining students were distributed between these two extremes, indicating substantial variation in problem-solving ability within the sample.

Table 11. Distribution of Student Measures

Measure (logit)	Students
+3	S12, S3, S37, S5
+2	S11, S13, S23
+1	S1
0 to +1	S10, S24, S34, S36, S57, S16, S2, S27, S29, S33, S4, S6, S22, S25, S62
0	S43, S45, S55, S56, S61, S66, S7, S74, S9
0 to -1	S17, S31, S44, S48, S49, S60, S65, S71, S41, S54, S63, S73
-1	S14
-1 to -2	S67, S69, S39, S40, S72, S42, S68
-2	S46, S50, S51, S52, S58, S64

The distribution of student measures also shows that student ability extended from approximately -2 to $+3$ logits, whereas the item locations were concentrated within a narrower range around the center of the scale. This indicates that the instrument targeted the central portion of the students' ability distribution more closely, while fewer items were located at the extreme upper and lower ends. Consequently, the precision of measurement may be lower for students whose ability estimates fall substantially above or below the main range of item locations. This pattern is particularly relevant for interpreting the instrument because the sample included students with varying levels of physics problem-solving ability. Overall, the Wright Map suggests that the instrument was able to represent a broad range of student ability, although additional items targeting the upper and lower ends of the ability continuum could improve measurement coverage in future applications.

4.5. Category Functioning

Category functioning was examined by inspecting the Rasch-Andrich thresholds for each item's response categories. A response category was considered to function appropriately when the thresholds progressed monotonically across successive categories. Five items (1a, 1b, 1c, 1d, and 2b) showed ordered thresholds, indicating that their response categories followed the expected progression along the ability continuum. In contrast, three items (2a, 2c, and 2d) exhibited disordered thresholds, indicating that the response categories did not progress monotonically across the latent ability scale. The threshold estimates and their standard errors are presented in Table 12.

Table 12. *Summary of Rasch-Andrich Thresholds per Item*

Item	Threshold 1		Threshold 2		Threshold 3		Status
	Measure	SE	Measure	SE	Measure	SE	
1a	-1.29	0.41	1.29	0.29	-	-	Ordered
1b	-2.24	0.45	2.24	0.36	-	-	Ordered
1c	-0.14	0.48	0.14	0.31	-	-	Ordered
1d	-0.73	0.42	0.73	0.29	-	-	Ordered
2a	-0.47	0.37	1.96	0.34	-1.49	0.34	Disordered
2b	-1.25	0.29	1.25	0.39	-	-	Ordered
2c	0.73	0.34	0.01	0.35	-0.73	0.34	Disordered
2d	0.59	0.31	-0.59	0.33	-	-	Disordered

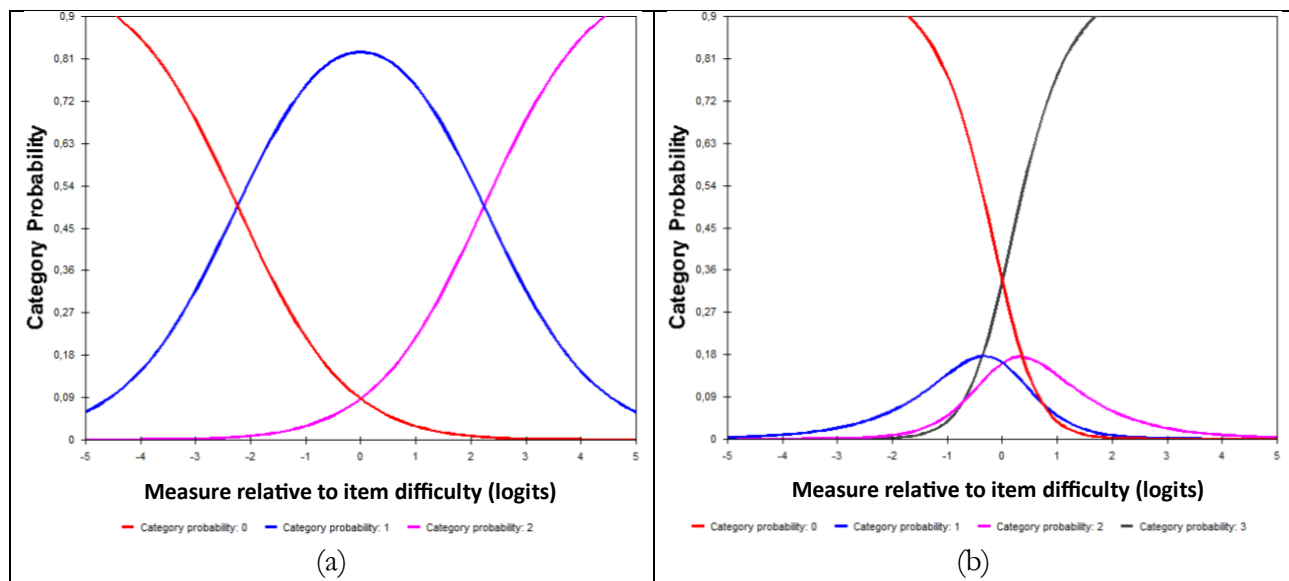


Figure 2. Category probability curves for (a) Item 1b (ordered thresholds) and (b) Item 2c (disordered thresholds).

For Item 1b, the thresholds were ordered at -2.24 and $+2.24$ logits, indicating a clear progression from Category 0 to Category 1 and then to Category 2 as student ability increased. As shown in Figure 2(a), each response category reached its highest probability within a distinct region of the measure continuum. Category 0 was most probable at lower ability levels, Category 1 dominated the middle range, reaching a peak probability of approximately 0.83 around a measure of 0, and Category 2 became most probable at higher ability levels. This pattern is consistent with the ordered threshold estimates.

In contrast, Item 2c showed disordered thresholds, with threshold estimates of 0.73, 0.01, and -0.73 logits. The thresholds did not progress in the expected increasing order, indicating that the intermediate response categories did not function as distinct steps along the ability continuum. As shown in Figure 2(b), Categories 0 and 3 dominated the lower and upper portions of the measure continuum,

respectively, whereas Categories 1 and 2 remained relatively unlikely across the continuum. Their curves did not form distinct regions in which either category became the most probable response. This pattern provides visual evidence consistent with the disordered threshold estimates.

Overall, five items demonstrated appropriate category functioning, whereas three items showed disordered thresholds. The disordered categories suggest that some intermediate scores may not have provided sufficient distinction between adjacent levels of students' problem-solving performance. This finding is particularly relevant for interpreting the functioning of the polytomous scoring structure and is considered further in the discussion.

4.6. Differential Item Functioning (DIF)

DIF analysis was conducted to determine whether item difficulty differed systematically between male and female students after accounting for differences in overall ability. Based on Draba's (1977) criteria, an item was considered to exhibit meaningful DIF when the absolute DIF contrast exceeded 0.5 logits and the absolute t-statistic was at least 1.96. As shown in Table 13, four of the eight items met both criteria and were therefore identified as exhibiting significant DIF.

Table 13. *Differential Item Functioning (DIF) Analysis by Gender*

Item	Item ID	Female	Male	DIF contrast	SE	t
1	1a	-0.51	-0.08	-0.43	0.47	-0.92
2	1b	-0.39	0.88	-1.27	0.58	-2.19
3	1c	-1.12	-1.22	0.10	0.48	0.22
4	1d	-0.88	-0.27	-0.61	0.44	-1.37
5	2a	-0.40	0.32	-0.72	0.30	-2.35
6	2b	1.15	0.79	0.37	0.47	0.79
7	2c	0.59	-0.13	0.72	0.30	2.44
8	2d	1.20	0.33	0.87	0.38	2.29

Note. t-values were computed as the ratio of the DIF contrast to its standard error; minor rounding discrepancies (≤ 0.04) between reported and recalculated t-values reflect rounding of the contrast and SE values to two decimal places.

Two items showed DIF favoring female students. Item 1b had a DIF contrast of -1.27 logits ($t = -2.19$), while Item 2a had a contrast of -0.72 logits ($t = -2.35$). The negative contrasts indicate that these items were estimated to be easier for female students than for male students after controlling for overall ability. Conversely, two items showed DIF favoring male students. Item 2c had a contrast of 0.72 logits ($t = 2.44$), and Item 2d had a contrast of 0.87 logits ($t = 2.29$), indicating that these items were more difficult for female students than for male students at comparable levels of overall ability.

The remaining four items did not meet both criteria for significant DIF. Item 1a had a contrast of -0.43 logits ($t = -0.92$), Item 1c had a contrast of 0.10 logits ($t = 0.22$), Item 1d had a contrast of -0.61 logits ($t = -1.37$), and Item 2b had a contrast of 0.37 logits ($t = 0.79$). Although Item 1d exceeded the 0.5 -logit magnitude criterion, its t-statistic did not reach the required level of statistical significance. It was therefore not classified as exhibiting significant DIF, although the relatively large contrast may warrant attention in future administrations.

Overall, the DIF analysis identified gender-related differences in four items, with two items favoring female students and two favoring male students. The presence of DIF in these items indicates that their difficulty estimates were not fully invariant across gender groups and should therefore be considered when interpreting item-level results.

4.7. Person Misfit

The person-fit analysis identified several students whose response patterns showed elevated Infit and/or Outfit MNSQ values (Table 14).

Table 14. *Students with Misfitting Response Patterns (Person Fit Statistics)*

Student	Measure (logit)	Infit MnSq	Outfit MnSq	Outfit ZStd
s63	-0.61	2.55	2.28	1.7
s24	0.94	2.27	1.81	1.3
s57	0.94	1.32	3.42	2.7
s66	-0.12	1.75	1.93	1.6
s44	-0.35	1.83	1.93	1.5
s55	0.08	1.46	1.99	1.7
s48	-0.35	1.23	1.71	1.2

Seven students (s24, s44, s48, s55, s57, s63, and s66) exhibited at least one MNSQ value above the criterion of 1.5. The most pronounced deviation was observed for s57, with an Outfit MNSQ of 3.42 and a ZSTD of 2.70. Students s63, s66, s44, and s24 also showed elevated Infit and/or Outfit MNSQ values, whereas s55 and s48 primarily showed elevated Outfit MNSQ values. Overall, the identified cases represent a small proportion of the 74 students, indicating that most students demonstrated response patterns that were reasonably consistent with the Rasch measurement model.

5. Discussion

5.1. Unidimensionality

The unidimensionality analysis indicated that the instrument consistently measured physics problem-solving ability, despite covering four Polya stages and two sound-wave sub-topics. The four stages, namely understanding the problem, devising a plan, carrying out the plan, and looking back, represent successive processes in solving a physics problem. Similarly, string frequency and sound intensity level provide different contexts in which students apply their problem-solving ability. Thus, the variation in problem-solving stages and content contexts did not result in separate measurement dimensions.

In physics education, this finding suggests that students' problem-solving ability in sound-wave problems can be considered an overall construct, even when the problems involve different stages of the problem-solving process and physical contexts. This indicates that the four Polya stages and the two sound-wave sub-topics are sufficiently connected to assess students' problem-solving ability as a whole rather than as separate constructs. Such an interpretation is consistent with previous studies showing that related sub-abilities can form a common underlying dimension when they are sufficiently connected (Uto, 2021; Wang & Bao, 2010).

5.2. Item and Person Reliability

Item Reliability. The high item reliability indicates that the relative ordering of item difficulties was estimated consistently in the present sample. This consistency can be related to the systematic development of the instrument. First, the items were constructed from a clearly defined problem-solving construct and were consistently organized according to Polya's four stages. Although the items differed in problem-solving stage and sound-wave context, each item remained within the same overall problem-solving framework. This consistency in construct specification reduced the likelihood that differences in item estimates were caused by substantial differences in what the items were intended to measure.

Second, the use of a structured scoring rubric provided consistent criteria for evaluating students' responses across the four Polya stages. The rubric specified the expected elements of students' problem-solving processes, allowing responses to be evaluated according to the same underlying construct even when the physical context or solution procedure differed. The expert validation process also contributed to this consistency by examining the alignment of the items with the intended construct, content, and language before empirical testing. The minor revisions identified during validation mainly concerned wording and language clarity, which helped reduce ambiguity in interpreting the tasks.

These features provide a plausible explanation for the stable item hierarchy observed in the Rasch analysis. Because the construct, problem-solving framework, scoring criteria, and item wording were systematically controlled during development, the resulting differences in item difficulty were more likely to reflect differences in the demands of the tasks rather than inconsistencies in item construction or scoring. However, this stability does not mean that item difficulty is determined solely by the Polya stage represented in each item. Difficulty may also be influenced by the physical context, mathematical relationships, amount of information, problem wording, and expected response. In the present instrument, the string-frequency and sound-intensity problems involved different physical relationships and types of mathematical reasoning. Therefore, the high item reliability should be interpreted as evidence of consistent item calibration within the developed instrument, rather than as evidence that a particular Polya stage is inherently more difficult than another. This interpretation is comparable with Mulyani et al. (2017), who reported stable item calibration within a single physics topic.

Person Reliability. In contrast, the person reliability indicates more limited precision in differentiating students with similar levels of physics problem-solving ability. This limitation may be related to the relatively small number of scored sub-items and the concentration of item difficulties within the moderate range. With fewer opportunities for students to demonstrate differences in ability, the instrument may be more effective for identifying broad differences in problem-solving ability than for making fine-grained distinctions between individual students.

This finding has implications for the further development of the instrument. A broader range of item difficulties would provide more opportunities to capture differences across the student ability continuum. The distribution of student ability relative to item difficulty in the Wright Map also indicates the need for additional items targeting ability levels that are less well represented. Therefore, future development should expand the number and difficulty range of sound-wave problems and involve larger, more heterogeneous samples to improve measurement precision.

5.3. Item Fit, Item Discrimination, and Item Difficulty Level

Item Fit. The item-fit results indicate that the response patterns were generally consistent with the Rasch model. However, the marginal fit of Item 1a suggests that this item may have differentiated students less effectively than the other items. This interpretation is supported by the stronger relationship observed for Item 2a, which assessed the same Polya stage. The difference between these items indicates that item functioning is not determined solely by the problem-solving stage being assessed. Instead, the physical information, mathematical relationships, wording, and response demands may influence how students respond to a particular task.

The relatively high ZSTD value of Item 2d also warrants attention, although the item remained within the specified criterion. Its response pattern may have been influenced by the more constrained response format compared with Item 1d, which required a more open-ended explanation. In physics problem solving, differences in response format may affect how clearly students' reasoning can be observed. Nevertheless, this possibility cannot be established from fit statistics alone and should be examined through students' responses and the characteristics of the items. The comparable standard errors across the items further suggest that the marginal fit of Item 1a was not associated with substantially lower estimation precision. Overall, Item 1a can be retained while receiving further review of its problem context, wording, and response structure. This interpretation is consistent with previous Rasch-based studies emphasizing that item-fit statistics should be considered together with the substantive characteristics of the items (Fisher, 2007; Mulyani et al., 2017; Oktaviani et al., 2026).

Item Discrimination. The positive Point Measure Correlation values indicate that all items contributed to the measurement of the intended physics problem-solving construct. This pattern is consistent with the structure of the instrument, in which the four Polya stages represent connected processes: interpreting a physical situation, developing a solution strategy, carrying out the procedure, and evaluating the result. Because these processes are related, students' performance on individual stages would be expected to show a positive relationship with their overall problem-solving ability.

The relatively weaker correlation of Item 1a, however, again suggests that the characteristics of a particular physics problem may affect how strongly an item represents the overall construct. Thus, the positive correlations should not be interpreted as evidence that all items function identically, but rather as an indication that the items generally contribute coherently to the assessment. This result is comparable with previous Rasch-based studies of physics problem-solving instruments, which also reported positive item–person relationships (Mulyani et al., 2017; Ridwan et al., 2023).

Item Difficulty Level. The items varied in difficulty, although most were classified as moderate. This variation provides different levels of challenge and indicates that the instrument was not concentrated at a single difficulty level. However, the difficulty pattern did not increase systematically across the Polya stages, since items representing the same stage could differ in difficulty. Therefore, item difficulty cannot be attributed solely to the problem-solving stage.

The easiest item assessed the carrying-out-the-plan stage within the string-frequency problem. This may indicate that the required procedure was relatively accessible once students had identified the relevant information and solution approach. In contrast, several items involving understanding the problem and devising a plan were more challenging. This pattern suggests that difficulties in sound-wave problem solving may arise before mathematical execution, particularly when students must identify relevant information and determine the appropriate physical relationship.

Nevertheless, the observed differences are likely influenced by the interaction between the problem-solving process and the conceptual and mathematical demands of each physics problem. The string-frequency and sound-intensity contexts require different physical relationships, information, and forms of mathematical reasoning. The predominance of moderately difficult items is comparable with Mulyani et al. (2017), while the variation from easy to difficult is consistent with Ridwan et al. (2023). The concentration of items at the moderate level nevertheless suggests that additional items with lower and higher difficulty could improve the coverage of the ability continuum.

5.4. Wright Map

The Wright Map indicates that the instrument represented variation in students' problem-solving ability, but the item difficulties were not distributed evenly across the ability continuum. Most items were concentrated around the central region, while some students were located beyond the range of available item difficulties. This suggests that the instrument was more closely targeted to students with moderate levels of problem-solving ability and provided less measurement information for students at the lower and higher ends of the continuum.

This targeting pattern may explain the relatively limited precision in distinguishing students with different ability levels. Students whose abilities were substantially higher or lower than the available item difficulties had fewer tasks through which differences in their problem-solving performance could be captured. This does not necessarily indicate that the items were inappropriate; rather, it suggests that the current item set provided less information for differentiating students at the extremes. The focused scope of the instrument may have contributed to this pattern, as the items were developed within the same

sound-wave content domain and based on the same Polya framework. Similar concerns regarding the importance of broad item-difficulty coverage have been reported in previous Rasch-based physics assessments (Ardianti et al., 2024; Maulana et al., 2026). Future development could therefore include sound-wave problems with lower and higher levels of cognitive demand to improve person–item targeting and provide more information across the ability continuum.

5.5. *Category Functioning*

The category functioning analysis indicates that the scoring structure did not function consistently across all open-ended physics problem-solving items. The presence of disordered thresholds suggests that some adjacent score categories were not sufficiently distinguishable as progressive levels of demonstrated problem-solving performance. Because the instrument uses polytomous scoring, this finding is important for determining whether higher scores meaningfully represent higher levels of students' problem-solving performance.

The concentration of disordered categories in the second problem may indicate that characteristics of the problem or its scoring criteria made some performance levels difficult to distinguish. This possibility is illustrated by the contrast between Item 1b, whose categories functioned in the expected order, and Item 2c, in which some middle categories did not emerge as distinct response levels. In the context of open-ended sound-wave problems, such a pattern may occur when adjacent scoring descriptors do not correspond to sufficiently different features of students' solutions. However, threshold statistics alone cannot determine the specific source of the disordering and should therefore be considered alongside students' responses and the scoring rubric.

From a physics-assessment perspective, the findings suggest that the scoring criteria for some sound-wave problem-solving responses require further examination. Descriptors for adjacent score levels could be clarified based on observable differences in students' physics reasoning, while categories that cannot be empirically distinguished may be reconsidered when supported by substantive evidence. In contrast, categories that function in the expected order can be retained. Thus, category functioning provides additional information for refining the scoring structure so that it more appropriately represents different levels of students' physics problem-solving performance.

5.6. *Differential Item Functioning (DIF)*

The DIF analysis indicates that several items functioned differently across gender groups even after accounting for differences in students' overall ability. The direction of DIF was balanced, with some items favoring female students and others favoring male students. This balanced pattern suggests that the DIF was not indicative of a general advantage for one gender across the instrument. Nevertheless, the presence of DIF means that performance on particular items may have been influenced by factors beyond the intended physics problem-solving construct.

The DIF items occurred across different sound-wave problems and stages of the problem-solving process. This pattern suggests that the observed DIF was not restricted to a particular Polya stage or content context. Since DIF statistics identify differences in item functioning but do not establish their causes, the observed differences cannot be attributed to gender itself. In the present instrument, factors such as item wording, physical context, information that must be interpreted, and response demands may interact differently with students from the two groups. This is particularly relevant to open-ended sound-wave problems, which require students to interpret physical information, select relevant relationships, and construct written solutions. Therefore, qualitative examination of the affected items and students'

responses is needed to determine whether particular item characteristics contributed to the observed DIF.

The remaining items did not show significant DIF, while Item 1d, despite showing a relatively larger contrast, did not meet the statistical significance criterion. This suggests that a relatively large group difference in an item is not sufficient on its own to classify the item as exhibiting DIF. The use of both the magnitude of the difference and its statistical significance is consistent with the approach described by Draba (1977). Overall, the flagged items should be reviewed particularly for wording, physical context, and response demands to determine whether the observed DIF reflects differences in item interpretation or unintended sources of item bias and to inform subsequent item revision.

5.7. Person Misfit

The person-fit analysis identified several students with response patterns that deviated from those expected by the Rasch model. These patterns indicate atypical responses rather than providing sufficient evidence that the students' responses were invalid. In open-ended physics problem solving, students may combine conceptual understanding, mathematical representation, procedural knowledge, and interpretation of physical situations in different ways. Consequently, a student's performance may not always follow the ability hierarchy expected across all problems.

Person misfit is particularly relevant to the present instrument because students were required to construct responses across different stages of solving sound-wave problems. A student may perform well in one stage but show less consistent performance in another, even when their overall ability is similar to that of other students. Person-fit analysis therefore provides additional information about individual response patterns that may not be apparent from total scores alone (Felt et al., 2017). This is consistent with Planinic et al. (2019), who highlighted the usefulness of Rasch analysis for identifying idiosyncratic response patterns in physics education research.

The relatively small number of students showing elevated person-fit statistics suggests that atypical response patterns were limited to particular cases rather than being a general characteristic of the sample. These cases nevertheless provide useful diagnostic information for examining students' responses and identifying whether particular sound-wave problems or problem-solving stages contributed to inconsistent performance. Thus, person-fit analysis complements item-level evaluation by providing an individual-level perspective on how students responded to the physics problem-solving tasks.

5.8. Strengths and Implications for Physics Education

This study contributes to the development of physics problem-solving assessment by providing a context-specific instrument based on Polya's four-stage framework. By representing problem solving through understanding the problem, devising a plan, carrying out the plan, and looking back, the instrument provides a basis for assessing students' problem-solving processes rather than relying solely on final-answer correctness. A further contribution lies in the comprehensive application of Rasch measurement, which examined not only item reliability but also person reliability and separation, item functioning, targeting, category functioning, DIF, and person fit. This approach provides diagnostic information that can guide further refinement of the instrument.

From a physics education perspective, the instrument can provide information about students' performance across different problem-solving stages, helping educators identify aspects of problem solving that may require greater instructional attention. For example, responses across Polya's stages may provide information about students' ability to interpret physical situations, develop solution strategies, carry out procedures, and evaluate their solutions. However, such stage-specific interpretations should

be considered alongside the overall measurement results rather than treated as independent measures of ability.

The findings also have implications for future physics assessment development. The moderate person reliability and limited targeting suggest the need for additional items covering a wider range of difficulty, while the disordered thresholds and DIF findings highlight the importance of reviewing scoring categories and item characteristics before broader application. Although the present instrument focuses on sound-wave problems, its Polya-based structure can be adapted to other physics topics and evaluated using Rasch analysis to examine whether similar measurement properties are obtained.

5.9. Limitations and Future Directions

Several limitations should be considered when interpreting the findings. First, the instrument was administered to 74 Grade 11 students from a single public high school, which may limit the generalizability of the estimated item and person parameters. Second, the instrument consisted of only eight scored sub-items, which contributed to moderate person reliability and separation and limited coverage of the lower and higher ends of the ability continuum. Future studies should therefore involve larger and more heterogeneous samples and include additional items with a wider range of difficulty levels.

In addition, three items showed disordered Rasch-Andrich thresholds, indicating that some adjacent scoring categories may not have represented clearly distinguishable levels of performance. Four items also exhibited gender-related DIF, although statistical DIF alone does not establish its underlying cause. The person-fit analysis further identified several students with atypical response patterns. These findings highlight the need to review the scoring rubric and DIF items qualitatively and to examine atypical responses more closely in future studies. Finally, because the instrument was developed for sound-wave problem solving, further research is needed to examine whether the Polya-based instrument structure demonstrates similar measurement properties when adapted to other physics topics.

6. Conclusion

Based on the results of the psychometric analysis using the Rasch model, the physics problem-solving ability instrument on sound wave material demonstrates good overall measurement quality. The instrument satisfies the assumption of unidimensionality, indicating that it consistently measures a single underlying problem-solving construct. The Rasch analysis showed that seven items met the specified empirical fit criteria, while Item 1a showed marginal fit and was retained based on its positive Pt Measure Corr and substantive relevance to the problem-solving construct. Overall, the instrument demonstrated acceptable measurement properties, although further refinement is recommended, particularly for items showing marginal fit, disordered response categories, and differential item functioning. The instrument shows a high level of item reliability alongside a fair level of person reliability, and item difficulty levels are distributed across easy, moderate, and difficult categories in line with Polya's problem-solving stages. All items also display positive discriminating power. Nevertheless, the analysis identified specific items requiring closer attention: one item showed a comparatively weaker relationship with the overall construct, several items exhibited disordered response-category thresholds, and about half of the items showed differential functioning between male and female students. These findings indicate that, while the instrument as a whole performs well, certain items would benefit from qualitative review to ensure they measure problem-solving ability equivalently across students.

Overall, this instrument is suitable for objectively measuring senior high school students' physics problem-solving ability on sound wave material, while also pointing to specific items that require refinement before wider use. Future development is recommended to review and revise the flagged items, reconsider the response category structure where needed, and involve a larger and more diverse student sample to strengthen the instrument's precision and generalizability

Authors Contribution

Rahma Alliya Aquilla: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Hera Novia:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Nurdini Nurdini:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Ethical statement

This study was conducted in accordance with applicable ethical principles for educational research. Prior to data collection, official research permission was obtained from the principal of SMA Negeri 1 Cimahi. All student participants were informed of the study's purpose and their voluntary participation before the test was administered. No sensitive personal data were collected, and all responses were anonymized and used solely for research purposes. As this study involved no clinical intervention, no deception, and no collection of sensitive personal information, formal review by an institutional ethics committee was not required.

Declaration of AI Use

The authors used GenAI to improve sentence clarity and readability. All AI-assisted results have been substantially reviewed, edited, and verified by the authors, who remain fully responsible for the accuracy, originality, scientific integrity, and the entire content of the final manuscript.

Conflict of Interest

The authors declare that there are no conflicts of interest, whether financial or non-financial, that could directly or indirectly influence the results of this study.

Acknowledgements

The researcher would like to express special thanks to the two academic advisors for their input, suggestions, and guidance throughout the process of writing this article, the physics teachers at Cimahi State High School 1, who were willing to set aside time for several meetings at the school for this research, as well as the students in classes XI A2 and XI A3 for the 2025–2026 school year, who devoted their time and thought to participating in this research.

References

- Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/10.1177/0013164485451012>
- Amiroh, D., Sibua, S., & Salim, A. (2021). Pendekatan Multi Representasi untuk Meningkatkan Penguasaan Konsep dan Pemecahan Masalah Mahasiswa pada Materi Gelombang. *BRILLIANT: Jurnal Riset Dan Konseptual*, 6(2), 290–302. <https://doi.org/10.28926/briliant.v6i2.615>
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Ardianti, R., Razak, A., Alberida, H., Fadilah, M., & Sari, M. (2024). *Analytical Thinking Skill and Problem Solving Skills Instruments in Biology Learning Using Rasch Model*. 8(4), 785–795.

<https://doi.org/10.1007/BF02293814>

- Boone, W., & Staver, J. (2020). *Advances in Rasch Analyses in the Human Sciences*.
- Docktor, J., Heller, K., & Heller, P. (2007). Robust assessment instrument for student problem solving. *AAPT Summer Conference* 2007, 2007. http://groups.physics.umn.edu/physed/People/Docktor/talks_papers/AAPT_S2007_abstract_p.pdf
- Docktor, J. L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K. A., Mason, A., Ryan, Q. X., & Yang, J. (2016). Assessing student written problem solutions : A problem-solving rubric with application to introductory physics. *Physical Review Physics Education Research*, 010130, 1–18. <https://doi.org/10.1103/PhysRevPhysEducRes.12.010130>
- Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. *Physical Review Special Topics - Physics Education Research*, 10(020119), 1–58. <https://doi.org/10.1103/PhysRevSTPER.10.020119>
- Dörner, D., & Funke, J. (2017). Complex Problem Solving : What It Is and What It Is Not. *Frontiers in Psychology*, 8(July), 1–11. <https://doi.org/10.3389/fpsyg.2017.01153>
- Draba, R. E. (1977). *The Identification and Interpretation of Item Bias*.
- Felt, J. M., Castaneda, R., Tiemensma, J., & Depaoli, S. (2017). Using person fit statistics to detect outliers in survey research. *Frontiers in Psychology*, 8, 863. <https://doi.org/10.3389/fpsyg.2017.00863>
- Fisher, W. P. J. (2007). Rating Scale Instrument Quality Criteria. *Rasch Measurement Transactions*, 21(1), 1087–1096. http://www.education.uwa.edu.au/Httpwww.Education.Uwa.Edu.Aunews/Rasch_Conference
- Frank, A., Akrosumah, S. M., Amenyek, F. B., & Sekyere, F. A. (2025). Enhancing Students ' Understanding of Sound and Waves through Multimedia-Supported Instruction : A Mixed-Methods Study in Ghana. *TJER*, 12(11), 499–510. https://www.researchgate.net/publication/397935462_Enhancing_Students'_Understanding_of_Sound_and_Waves_through_Multimedia-Supported_Instruction_A_Mixed-Methods_Study_in_Ghana
- Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-todd, J., Foltz, P. W., & Hesse, F. W. (2018). *Advancing the Science of Collaborative Problem Solving*. <https://doi.org/10.1177/1529100618808244>
- Heller, P., & Holla Baugh, M. (1992). Teaching Problem Solving through Cooperative Grouping. Part 2: Designing Problems and Structuring Groups. *American Journal of Physics*, 60(637). <https://doi.org/10.1119/1.17118>
- Iribarra, D. T. (2018). Rasch Model. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, 4, 1371–1374. <https://doi.org/10.4135/9781506326139.n574>
- Lestari, P. E., Purwanto, A., & Sakti, I. (2019). Pengembangan instrumen tes keterampilan pemecahan masalah pada konsep usaha dan energi di SMA. 2(3), 161–168. <https://doi.org/10.33369/jkf.2.3.161-168>
- Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, 16(2), 878. https://www.researchgate.net/publication/305377322_What_do_infit_and_outfit_mean-square_and_standardized_mean
- Listiawan, N. A., Aviyanti, L., & Arifiyanti, F. (2026). Analisis Instrumen Kemampuan Pemecahan Masalah Fisika Pada Materi Kalor Menggunakan Pendekatan Model Rasch. *Jurnal Pendidikan, Sains, Geologi, Dan Geofisika*, 7(2). <https://doi.org/10.29303/goescienceed.v7i2.1928>
- Maloney, D. P. (2011). An Overview of Physics Education Research on Problem Solving Abstract : *Getting Started in Physics Education Research*, 2, 1–33. <https://www.compadre.org/repository/document/ServeFile.cfm?ID=11457>

- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47, 149–174. <https://doi.org/10.1007/BF02296272>
- Maulana, I., Syarah, E., & Nainggolan, E. (2026). *Validation of Instrument to Measure Pre-Service Science Teachers Digital Skills : Rasch Model*. 11, 1987–1996. <https://doi.org/10.29303/jipp.v11i2.4794>
- Mukhopadhyay, D. R. (2013). Problem Solving In Science Learning - Some Important Considerations of a Teacher. *IOSR Journal of Humanities and Social Science*, 8(6), 21–25. <https://doi.org/10.9790/0837-0862125>
- Mulyani, S., Efendi, R., & Ramalis, T. R. (2017). *Karakterisasi tes keterampilan pemecahan masalah fisika berdasarkan teori respons butir*. <https://doi.org/10.52434/jpif.v1i1.1006>
- Musengimana, T., Yadav, L. L., Uwamahoro, J., & Nizeyimana, G. (2025). Assessing physics students' problem-solving skills : a baseline investigation. *Discover Education*, 4(196), 1–16. <https://doi.org/10.1007/s44217-025-00640-1>
- Nurrohman, I., Aviyanti, L., & Kaniawati, I. (2026). Instrumen Kemampuan Creative Problem Solving Siswa SMA pada Materi Energi Alternatif : Analisis Model Rasch. *Jurnal Pendidikan, Sains, Gelologi, Dan Geofisika*, 7(2). <https://doi.org/10.29303/goescienceed.v7i2.1840>
- Oktaviani, I., Lestari, P. A., Isnanto, E. P., Putri, K. N., Ardiansyah, M. I., & Suwanto, S. (2026). Analisis Kualitas Butir Soal Tes Biologi Menggunakan Rasch Model. *Jurnal Pendidikan Ilmu Pengetahuan Alam (JP-IPA)*, 07(02), 219–228. <https://doi.org/10.56842/jp-ipa.v7i2.1002>
- Özreçberoğlu, N., & Çağanaga, Ç. K. (2018). Making It Count : Strategies for Improving Problem-Solving Skills in Mathematics for Students and Teachers' Classroom Management. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(4), 1253–1261. <https://doi.org/10.29333/ejmste/82536>
- Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, 15(2), 020111. <https://doi.org/10.1103/PhysRevPhysEducRes.15.020111>
- Polya, G. (1957). *How to Solve it*. Doubleday Anchor Books.
- Putra, R. N., Usman, U., Wahyuni, A. S. A., Sakona, A. A. R., & Setiawan, T. (2025). Analisis Kemampuan Pemecahan Masalah Fisika Peserta Didik Kelas XI Fisika SMA Negeri 2 Makassar. *AL-IRSYAD Journal of Physics Education*, 4, 99–110. <https://doi.org/10.58917/ijpe.v4i2.403>
- Ramadhani, A. S., & Handayani, I. (2025). *Analisis Kualitas Instrumen Minat Belajar dalam Pembelajaran Matematika: Pendekatan Model Rasch*. 17, 353–367. <https://doi.org/10.26618/sigma.v17i1.18344>
- Ridwan, I. M., Kaniawati, I., Suhandi, A., & Ramlan, T. (2023). Instrumen Kemampuan Pemecahan Masalah pada Tema Perubahan Iklim : Analisis Rasch Model. *DIFFRACTION: Journal for Physics Education and Applied Physics*, 5(1), 37–46. <https://doi.org/10.37058/diffraction.v5i1.7738>
- Rokhmat, S., & Hanif, A. (2017). Pengembangan Instrumen Tes Keterampilan Pemecahan Masalah pada Materi Getaran, Gelombang, dan Bunyi. *JPPPF (Jurnal Penelitian dan Pengembangan Pendidikan Fisika)*, 3, 157–166. <https://doi.org/10.21009/1.03206>
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Asessment Pendidikan*. Trim Komunikata. <https://jbse.ulm.ac.id/index.php/PMPIPA/article/view/22/0>
- Susdelina, S., Perdana, S. A., & Febrian, F. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik dan Rasch Model. *Jurnah Kiprah*, 1, 41–48. <https://doi.org/10.31629/kiprah.v6i1.574>
- Treffinger, D. J., & Isaksen, S. G. (2005). The History , Development , and Implications for Gifted Education and Talent Development Creative Problem Solving: *Gifted Child Quarterly*, 49(4), 342–353. <https://doi.org/10.1177/001698620504900407>
- Uto, M. (2021). A multidimensional generalized many - facet Rasch model for rubric - based performance

assessment. *Behaviormetrika*, 48(2), 425–457. <https://doi.org/10.1007/s41237-021-00144-w>

Wang, J., & Bao, L. (2010). Analyzing Force Concept Inventory with Item Response Theory. *American Journal of Physics*, 78(10), 1–18. <https://doi.org/10.1119/1.3443565>

Appendix A. Problem-Solving Skills Assessment Rubric

Problem-Solving Skills Indicators	Description	Score
Understanding the Problem	Writes down the problem and all physical quantities in the question using the correct units	3
	Write the problem and/or the physical quantities in the question using incorrect units or without units	2
	Stating the problem in an illogical manner and/or identifying only one physical quantity	1
	Failing to identify a physical quantity and failing to state the problem	0
Planning a Strategy	Writing down concepts and planning the steps to be taken completely and accurately	2
	Stating concepts and/or planning steps only, but not very accurately	1
	Not planning anything and not writing down any ideas	0
Problem Solving	Solve problems correctly and include the units	3
	Solving problems correctly but not including the units	2
	Solves the problem somewhat incorrectly or does not complete it	1
	Did not solve the problem	0
Reviewing the results	Writing accurate and reasonable conclusions	2
	Drawing conclusions that are inaccurate or illogical	1
	Failing to write a conclusion	0
Maximum Score		10

Appendix B. Problem-Solving Instrument Items

Note: The instrument was originally administered to students in Indonesian; the English version below is a translation for reporting purposes.

Item 1 (1a–1d): String Frequency Problem

A musical instrument technician is tuning a grand piano 30 minutes before a solo concert begins. Suddenly, the steel wire for note E (frequency 330 Hz) snaps because it was pulled too tight. The problem is that the technician forgot to bring a spare wire for note E. In his equipment bag, he only finds a spare wire for note A (frequency 440 Hz). Since time is very short and it is impossible to find a replacement wire outside the venue, the technician must find a way to install the note-A wire so that it produces the sound of note E instead. This spare note-A wire happens to have exactly the same length and thickness (mass per unit length) as the wire that broke. According to the manufacturer's specifications, this note-A wire must normally be pulled with a tension force of 800 N to produce exactly 440 Hz.

Item	Polya Stage	Question
1a	Understanding the Problem	Based on the description above, what problem is the technician facing? State the physical quantities that are already known and the quantity that must be determined.
1b	Devising a Plan	What physics concept applies to the problem above? Plan the steps you would take to help the technician solve the problem.
1c	Carrying Out the Plan	Solve the problem based on the steps you planned in point (b).
1d	Looking Back	Is the quantity from your calculation reasonable and consistent with the applicable physics concept? What action should the technician take on the wire-tuning mechanism?

Item 2 (2a–2d): Sound Intensity Level Problem

A music festival committee is testing the main-stage speaker tower on an open, flat field. When tested at maximum volume, the committee's sound meter, placed 10 meters directly in front of the speaker, recorded an extraordinary sound intensity level (SIL) of 140 dB. According to the event schedule, the DJ's peak performance is set to run continuously at that maximum volume for 15 minutes. According to WHO hearing-safety standards, the maximum safe exposure time is inversely related to the sound intensity level, as follows: 100 dB $\rightarrow$ 15 minutes; 114 dB $\rightarrow$ 30 seconds; 120 dB $\rightarrow$ 1 second; >120 dB $\rightarrow$ 0 seconds. The committee must determine a safe distance for installing the barrier fence (separating the VIP audience) so the audience's hearing is not damaged after the 15-minute peak performance. Three barrier-placement options are available: Option A: 100 m; Option B: 200 m; Option C: 1,000 m (1 km). (Logarithm aid: $\log 10 = 1$; $\log 20 = 1.3$; $\log 100 = 2$.)

Item	Polya Stage	Question
2a	Understanding the Problem	Based on the description above, what problem must the committee solve? State the physical quantities and safety-limit information that are known, and what is being asked.
2b	Devising a Plan	What physics concept applies to the problem above? Plan the systematic steps you would take to help the committee solve the problem.
2c	Carrying Out the Plan	Solve the problem by carrying out the mathematical calculation according to the steps you planned in point (b).
2d	Looking Back	Compare the numbers from your calculation with the WHO guideline table above. Which location option would you ultimately recommend to the committee, and explain your scientific reasoning for why the other options are dangerous for the audience.