JPPIPA 11(12) (2025)

Jurnal Penelitian Pendidikan IPA

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

Integrating Technology and Psychometrics: Developing an App-Based Diagnostic Tool Using the Partial Credit Model to Uncover Student Misconceptions

Nurhasanah[1*], Zul Hidayatullah[2], Hajra Yansa[3], Moh. Badrus Sholeh Arif[1], Muh Asriadi[4]

[1] Primary Teacher Education, Faculty of Education, Universitas Jember, Jawa Timur, Indonesia.
[2] Science Education, Faculty of Education, Universitas Hamzanwadi, Nusa Tenggara Barat, Indonesia.
[3] Early Childhood Teacher Education, Faculty of Education, Universitas Musamus, Papua, Indonesia.
[4] Early Childhood Teacher Education, Universitas Pendidikan Indonesia, Jawa Barat, Indonesia.

Received: July 25, 2025
Revised: October 29, 2025
Accepted: December 25, 2025
Published: December 31, 2025

Corresponding Author:
Nurhasanah
nurhasanah.fkip@mail.unej.ac.id

DOI: 10.29303/jppipa.v11i12.12305

Abstract: Misconceptions in science learning remain a major barrier to students' conceptual understanding, while traditional assessments often fail to detect them effectively. This study aimed to develop *PhyTestApp*, an app-based diagnostic tool that integrates the Partial Credit Model (PCM) under Item Response Theory (IRT) to uncover student misconceptions in science. A research and development design was employed, involving expert validation, limited trials, and psychometric testing. The two-tier items were designed to capture both factual knowledge and reasoning. Findings indicated that the instrument met psychometric requirements, with items demonstrating good fit and functioning across different levels of student understanding. Usability testing also showed positive responses from students and teachers regarding clarity, content relevance, and technical operation. Overall, *PhyTestApp* provides a reliable and practical diagnostic tool that facilitates immediate feedback and supports more targeted science instruction. These results highlight the potential of combining psychometric modeling with mobile technology to improve the quality of science education and more effectively address misconceptions in line with 21st-century learning goals.

Keywords: Diagnostic assessment; Educational technology; Misconceptions; Partial credit model

Introduction

Science concepts are essential for explaining and understanding natural phenomena across scientific disciplines. They provide a practical framework for integrating different areas of science and play a crucial role in shaping how learners think and model both natural and technological processes (Soehato & Csapó, 2021). A solid conceptual understanding is therefore considered a key indicator of successful learning in science education. However, achieving this understanding is not straightforward. Teaching science requires addressing and reshaping students' pre-existing ideas or naïve theories about the world, which often diverge from scientifically accepted views (Theobald & Brod, 2021). In this regard, one of the most persistent challenges in science education is the presence of misconceptions—alternative conceptions that interfere with the accurate understanding of scientific knowledge (Rohmah et al., 2018).

Misconceptions differ fundamentally from mere gaps in knowledge or accidental mistakes. They represent structured beliefs that directly contradict widely accepted scientific explanations and often persist despite instruction (Çelikkanlı & Kızılcık, 2022). For example, students may consistently misinterpret force and motion as being synonymous, or they may assume that heavier objects fall faster than lighter ones, even after repeated instruction. These are not simply slips of memory, but entrenched mental models that conflict

How to Cite:
Nurhasanah, Hidayatullah, Z., Yansa, H., Arif, M. B. S., & Asriadi, M. (2025). Integrating Technology and Psychometrics: Developing an App-Based Diagnostic Tool Using the Partial Credit Model to Uncover Student Misconceptions. *Jurnal Penelitian Pendidikan IPA*, *11*(12), 861-872. https://doi.org/10.29303/jppipa.v11i12.12305

with scientific consensus (Dellantonio & Pastore, 2021). The persistence of such misconceptions highlights the difficulty of fostering conceptual change in science learning.

Research since the 1970s, influenced by Piagetian and Vygotskian perspectives, has emphasized that students bring intuitive frameworks to the classroom, which shape how they interpret instruction (Samsudin et al., 2024). Misconceptions—also referred to as preconceptions, alternative conceptions, or naïve theories—are increasingly seen as integral to the process of knowledge construction rather than as mere errors to be eliminated. Instructional approaches such as cognitive conflict and conceptual change strategies attempt to leverage these misconceptions as starting points for deeper learning, by challenging students' reasoning and fostering epistemic awareness (Vosniadou, 2020). Thus, identifying and understanding students' misconceptions is a crucial prerequisite for designing effective instruction.

The sources of misconceptions are diverse. They may arise from students' incomplete comprehension of classroom instruction, misinterpretations of abstract representations, reliance on everyday experiences, or even from teachers' own inaccurate explanations (Wahidah & Saptono, 2019). Misconceptions can also be reinforced by textbooks, analogies, or media representations that oversimplify or distort scientific concepts (Nainggolan et al., 2023). In physics education, where many concepts are abstract and counterintuitive, misconceptions are especially prevalent and often become major obstacles to consistent learning progress (Pujayanto et al., 2018). Left unaddressed, these misconceptions not only hinder the understanding of current material but may also propagate into subsequent topics, undermining long-term conceptual development (Juita et al., 2023; Wahyuningsih et al., 2013).

Given their persistence, misconceptions need to be systematically identified and diagnosed. Diagnostic tests are among the most widely used instruments for this purpose. They are designed not merely to evaluate knowledge acquisition but to reveal specific patterns of misunderstanding among learners (Suban et al., 2024). Administered before, during, or after instruction, diagnostic tests allow teachers to uncover what students know, partially know, and misunderstand (Taslidere, 2016; Tsui & Treagust, 2010). The value of diagnostic tests lies in their ability to capture students' reasoning processes, thereby enabling teachers to design more targeted interventions (Nainggolan et al., 2023). By revealing error patterns through systematic analysis, diagnostic tests can serve as an essential foundation for improving the quality of instruction and addressing learning difficulties at their root (Juita et al., 2023).

Over the years, various forms of diagnostic tests have been developed in science education. Among these, the two-tier diagnostic test has gained prominence. This format requires students to select an answer to a multiple-choice question (first tier) and then justify their choice or provide reasoning (second tier), making it particularly effective in distinguishing between correct answers based on guessing and genuine conceptual understanding (Chen et al., 2003; Gurel et al., 2015). The two-tier format is both student-friendly and teacher-friendly: it reduces random guessing, captures reasoning patterns, and provides clearer insights into students' misconceptions compared to traditional multiple-choice tests (Laliyo et al., 2019). Such tests are now widely recommended for diagnosing misconceptions in various science domains, including physics.

However, the effectiveness of diagnostic instruments depends on their psychometric quality. A diagnostic test must demonstrate validity, reliability, and appropriate difficulty levels to provide accurate results (Nurhasanah et al., 2024). In Indonesia and many other contexts, test development and analysis have traditionally relied on Classical Test Theory (CTT). While useful, CTT has notable limitations: item characteristics depend on the specific sample of test-takers, difficulty parameters are influenced by students' overall abilities, and scores reflect not only student ability but also the test design itself (Abidin, 2018). These weaknesses limit the precision of diagnostic assessments, particularly when the goal is to identify nuanced patterns of misconceptions.

To address these limitations, Item Response Theory (IRT) has been increasingly applied in educational measurement. Unlike CTT, IRT analyzes each test item independently of the test-taker population, enabling more precise estimation of both item parameters (such as difficulty) and student ability levels (Syamsuddin, 2023). IRT therefore provides a stronger methodological foundation for developing diagnostic instruments. Within IRT, different models can be used depending on the type of item. For multiple-choice two-tier diagnostic tests, scoring models such as the Partial Credit Model (PCM), the Graded Response Model (GRM), and the Modified Graded Response Model (MGRM) are commonly applied (Abidin, 2018). Among these, PCM is particularly suitable for items that involve multi-step reasoning or allow partial correctness, as it generates composite scores that reflect the depth of students' understanding (Istiyono, 2018). This makes PCM a robust approach for analyzing diagnostic tests aimed at detecting misconceptions in science learning.

At the same time, advances in educational technology have opened new opportunities for delivering diagnostic assessments in digital formats. The

rapid development of digital tools and mobile applications has transformed learning environments, making them more interactive, efficient, and accessible (Sukatiman et al., 2024). Paper-based assessments, while still common, are increasingly viewed as less practical due to their higher costs, delayed feedback, and limited adaptability to diverse student needs (Greiff et al., 2016; Salma, 2015). In contrast, application-based diagnostic tests offer several advantages: they streamline administration, reduce logistical burdens, enable real-time data analysis, and enhance student engagement (Sahidu et al., 2017). Moreover, digital platforms can seamlessly integrate advanced measurement models such as IRT, allowing for more sophisticated and precise analysis of student responses.

The integration of diagnostic testing, IRT analysis, and mobile technology thus represents a promising direction for addressing misconceptions in science education. Application-based diagnostic instruments have the potential to make misconception detection more accurate, accessible, and scalable, while also providing immediate feedback to both students and teachers. In particular, the use of the Partial Credit Model within such applications allows for nuanced measurement of students' conceptual understanding and partial reasoning, offering insights that traditional scoring models might overlook.

Based on these considerations, the present study aims to design and develop a diagnostic test in the form of a mobile application that applies the Partial Credit Model approach to detect misconceptions in science learning. By combining robust psychometric analysis with technology-enhanced assessment, this study seeks to contribute to both the methodological advancement of diagnostic testing and the practical improvement of science education, ensuring that misconceptions can be more effectively identified and addressed.

Method

Research Design

This study used a developmental research design to produce an application-based diagnostic test (PhyTestApp) for identifying students' misconceptions in physics. The workflow integrated the 4-D model (Define, Design, Develop, Disseminate; Thiagarajan, 1974) as the macro phases, with test-development substeps from Oriondo et al. (1984) nested within each phase.

In the Define phase, the team specified learning objectives, described the diagnostic construct, identified physics concepts prone to misconceptions, and prepared the test blueprint/table of specifications. The Design phase covered item writing for two-tier multiple-choice questions, development of the scoring scheme aligned with the Partial Credit Model (PCM), and prototyping of the application's user interface; expert content review was planned in this phase. The Develop phase conducted expert validation, a limited pilot, psychometric analysis under IRT/PCM, iterative item revision, assembly of the 11-item final form, and usability/feasibility checks. The Disseminate phase packaged the validated instrument as PhyTestApp and prepared it for broader classroom implementation.

Participant and Sample

The pilot involved 269 students selected via purposive sampling to match the target educational level and learning context. Expert validators comprised physics education specialists, experienced teachers, and assessment reviewers who provided judgments on content relevance and clarity. Student and teacher users also contributed usability and feasibility feedback during product evaluation.

Research Instruments

The diagnostic instrument consisted of two-tier multiple-choice items: five answer options (Tier-1) paired with five reasoning options (Tier-2). Scoring followed an ordered, partial-credit scheme that captured the consistency between selected answers and reasons, operationalized for PCM. Non-test instruments included structured validation sheets for experts and user questionnaires for feasibility/usability (content relevance, clarity, functionality). Content validity was summarized using Aiken's V based on expert ratings.

Procedures

After blueprinting and item writing, expert validation was conducted and results informed item revision. The revised instrument was piloted using the app environment. Response data were extracted for psychometric analysis, followed by iterative refinement and final test assembly (11 items). Usability and feasibility evidence was gathered from students and teachers to ensure practicality prior to wider deployment. The finalized instrument was then packaged and disseminated as PhyTestApp for broader classroom use.

Data Analysis

Psychometric analyses used the Partial Credit Model (PCM) within the Item Response Theory (IRT) framework. Analyses examined model–data fit (e.g., item fit indices), category functioning and threshold ordering, item difficulty parameters, and reliability/test information to judge measurement quality and coverage across ability levels. QUEST and PARSCALE software supported estimation and diagnostics.

Result and Discussion

Overview of Product Development and Implementation

The final product of this research is a diagnostic test instrument designed to identify students' misconceptions, specifically in the topic of sound and light waves. The test consists of 11 two-tier items, where the first tier presents a multiple-choice question and the second tier provides five reasoning options. Following expert validation and field trials, the instrument was deemed both valid and reliable, and subsequently integrated into a digital platform named PhyTestApp (Physics Test Application). The application was administered to 11th-grade students after instruction on the target topics was completed.

The test items were developed following the instrument development model proposed by Oriondo et al. (1984), while the digital platform was developed using the 4D Model consist Define, Design, Develop, and Disseminate (Thiagarajan, 1974). The application was built using a cross-platform programming language to support deployment on both web and mobile devices, enabling flexible, online access.

Definition Stage

This stage began with the identification of assessment challenges, informed by a comprehensive literature review. The review focused on learning difficulties in physics, 21st-century skills, and innovations in technology-based assessments. Insights from the literature were used to conceptualize a diagnostic instrument aimed at effectively detecting students' conceptual misunderstandings. Several limitations of traditional paper-based assessments were identified, including lack of efficiency, delayed feedback, and difficulty in detecting misconceptions in real time. To overcome these issues, PhyTestApp was designed not only to assess misconceptions but also to support learning by integrating diagnostic feedback, tailored recommendations, and supplementary learning resources in the form of videos and PDF materials. In contrast to conventional tests, PhyTestApp provides immediate and actionable feedback aligned with students' responses, thereby facilitating more meaningful and personalized learning experiences.

Media and Test Design Stage

The quality of diagnostic test items is critically influenced by three key aspects: content accuracy, construct alignment, and linguistic clarity. These elements were operationalized through a detailed test item specification table, which served as a blueprint for guiding item development and ensuring item feasibility. Following the established specifications, a total of 11 two-tier diagnostic items were developed to detect students' misconceptions. Each item was carefully designed not only to assess conceptual understanding but also to provide diagnostic feedback and individualized recommendations based on various student response patterns. This feature enhances the instrument's function beyond simple assessment, allowing it to serve as a formative learning tool.

An illustration of a representative item from the test is presented in Figure 1, demonstrating the integration of tiered responses and associated feedback for misconception diagnosis.

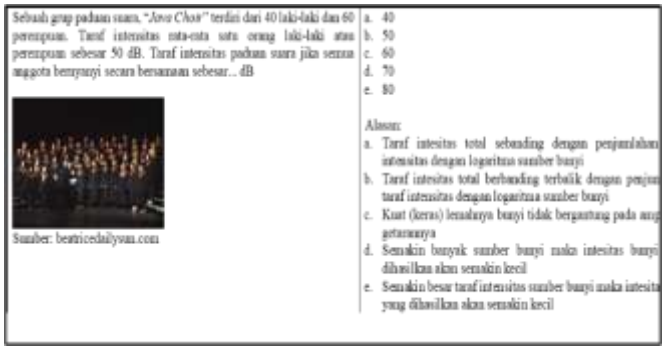

Figure 1. Example of a two-tier diagnostic test item designed to identify students' misconceptions

Development, Test Trial, and Test Assembly Stage

Following the completion of the media design, the development phase of the PhyTestApp was carried out. The application was built using a mobile-friendly platform and subsequently published on the Google Play Store, allowing students to easily download and access the app using Android devices with an internet connection. The interface and main features of the PhyTestApp are illustrated in Figure 2.

Figure 2. User interface and key features of the phytestapp mobile application

Figure 3 illustrates the web-based interface of the PhyTestApp platform, a core component of its digital ecosystem. The PhyTestApp ecosystem includes a web-based content management system that supports two types of users: administrators and teachers. The administrator role, typically handled by the system developer, has full access to all content and

functionalities within the platform. In contrast, teachers are granted access to features that support instructional use, such as managing student groups, distributing diagnostic test materials, and retrieving diagnostic reports.

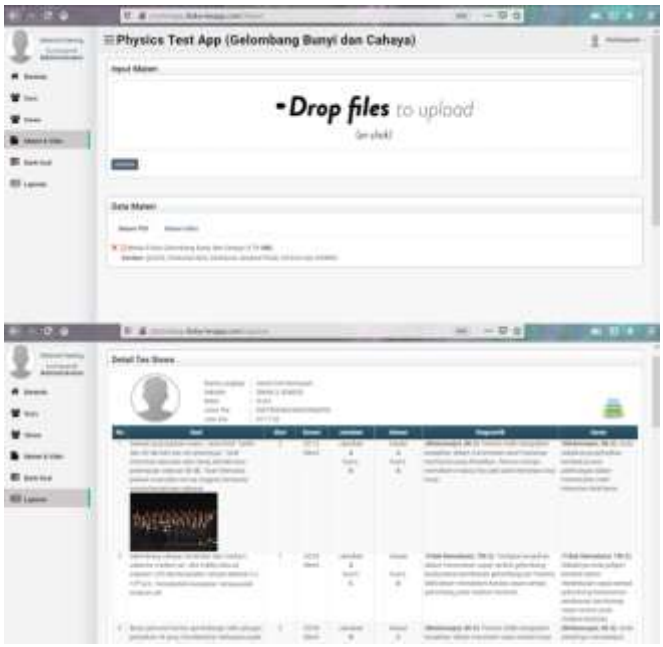

Figure 3. Web-based interface of the PhyTestApp platform used for test management and result monitoring

Through this interface, teachers can access both individual and class-level performance data. The ability to download and analyze student results facilitates formative assessment and allows for data-driven instructional planning. Additionally, the platform enables educators to monitor responses in real time, helping them quickly identify students' misconceptions and target specific areas for remediation.

The interface is designed to be user-friendly and accessible, ensuring usability across varying levels of digital literacy among educators. Its clear layout and intuitive navigation enhance the practicality of the system in real educational settings. Overall, this web-based platform not only streamlines test administration and monitoring but also empowers teachers to make informed pedagogical decisions based on reliable diagnostic data.

Development Results of the Diagnostic Test Instrument

The development of the misconception diagnostic test involved expert validation based on three critical dimensions: content relevance, item construction, and language clarity. To evaluate the content validity, a structured validation sheet was distributed to eight expert validators, each of whom assessed all 11 test items. Content validity refers to the extent to which the items accurately reflect the intended constructs and instructional content. In this study, Aiken's V coefficient was employed to quantify the degree of expert agreement for each item, using a scale that indicates item validity levels. All 11 two-tier diagnostic test items were evaluated as valid, with Aiken's V values exceeding the acceptable threshold. The detailed Aiken's V results for each item are summarized in Table 1.

Table 1. Aiken's V Content Validity Coefficients for the Diagnostic Test Items

Items	Validity score	Criteria
1.	1	Highly valid
2.	0.94	Highly valid
3.	0.94	Highly valid
4.	0.97	Highly valid
5.	0.94	Highly valid
6.	1	Highly valid
7.	0.97	Highly valid
8.	1	Highly valid
9.	1	Highly valid
10.	1	Highly valid
11.	1	Highly valid

As shown in Table 1, the Aiken's V coefficients for all items ranged from 0.94 to 1.00, exceeding the minimum threshold of 0.75 for 8 raters with 4-point scales (Aiken, 1985). According to Istiyono (2018), items with $V > 0.8$ are considered to have high content validity. Therefore, all 11 diagnostic test items were deemed valid.

In addition to numerical validation, qualitative feedback from validators emphasized the need for stronger alignment with cognitive indicators, greater linguistic precision, and consistency in symbols and notations. Several items required revisions, such as refining wording for clarity, correcting the use of vector symbols and units, and standardizing mathematical and scientific notations. Validators also recommended the inclusion of relevant visual elements to aid comprehension. These revisions were incorporated prior to empirical testing to ensure that all items met the expected standards of clarity, alignment, and construct relevance.

Media Feasibility Test Results

The feasibility of the PhyTestApp was evaluated through expert review using a structured questionnaire that assessed two key aspects: appearance (design) and effectiveness of use. The results are presented in figure 4.

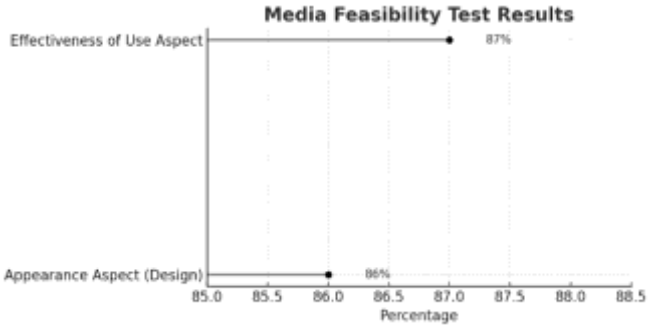

Figure 4. Media feasibility test results

Based on Akbar (2013), both assessed aspects fall into the highly feasible category. As shown in Figure 4, the appearance and effectiveness of use received similarly high scores, reflecting the media's strong visual appeal and functional quality. These findings suggest that PhyTestApp is both aesthetically acceptable and pedagogically effective, making it suitable for classroom implementation. The close range of scores further indicates consistent quality across its visual and usability dimensions.

In addition to quantitative ratings, qualitative feedback was also obtained from expert validators. They suggested several improvements to enhance the user experience and technical quality of the application. First, the media background was considered too minimalistic, and experts recommended adopting a design that is visually appealing yet still simple. They also advised including an introductory explanation on the post-login homepage to inform users about the application's purpose, target grade level, and version information. Furthermore, it was recommended that the registration system be made more flexible and automated to increase accessibility for both teachers and students. These suggestions were used as the basis for further refinement of the application prior to large-scale implementation.

Results of the Limited Trial of Diagnostic Misconception Items

A limited trial was conducted as a preliminary step to ensure the quality of the diagnostic misconception items prior to full-scale implementation. The primary objectives of this stage were to (1) identify items that were suitable for further use, (2) detect items requiring revision, and (3) eliminate items that failed to meet psychometric standards. The item analysis process employed the Partial Credit Model (PCM) within the framework of Item Response Theory (IRT). The characteristics assessed in this analysis included item fit, score reliability, Item difficulty levels, Item Characteristic Curves (ICCs), and the Test Information Function. These analyses were used to guide item revisions and finalize the instrument before integration into the PhyTestApp.

Item Fit (Goodness of Fit)

The empirical validity of each test item was evaluated through a goodness-of-fit analysis based on the Partial Credit Model (PCM). This analysis aimed to determine how well individual items aligned with the expectations of the PCM, thereby indicating the extent to which students' responses were consistent with their ability levels in relation to item difficulty.

Table 2. Estimation Results of the Diagnostic Misconception Test Instrument

Item parameters	Item estimation	Testee estimation
Mean and standard deviation of infit mnsq	0.99 ± 0.15	1.00 ± 0.39
Mean and standard deviation of infit t	0.22 ± 2.09	0.01 ± 1.01
Items or testees with a score of 0	0	0
Items or testees with a perfect score	0	0
Reliability estimation	0.71	0.77

The item fit was assessed using the INFIT Mean Square (INFIT MNSQ) and its standardized value (INFIT t), generated by the Quest software. The INFIT MNSQ statistic evaluates the consistency of responses from test-takers whose abilities are close to an item's difficulty level, with values near 1.0 indicating a good fit, while values far above or below 1 suggest potential misfit. According to measurement standards (Bond, 2015), INFIT MNSQ values between 0.77 and 1.30 are generally acceptable, and INFIT t values are expected to fall within the range of –2 to +2 to indicate a satisfactory fit. The results of the item fit analysis for the Misconception Diagnostic Instrument are presented in Table 2.

Based on the empirical data from the limited trial, analyzed using the Quest program, the goodness-of-fit results were obtained for each item and for the test as a whole. The fit analysis aims to determine whether each item aligns well with the expectations of the Partial Credit Model (PCM) under the 1-Parameter Logistic (1PL) framework.

The item fit was assessed using the mean INFIT Mean Square (INFIT MNSQ) value and its standard deviation, as well as the standardized INFIT t value. As shown in Table 2, the mean INFIT MNSQ for the developed instrument was 0.99, with a standard deviation of 0.15, indicating an overall good fit to the PCM.

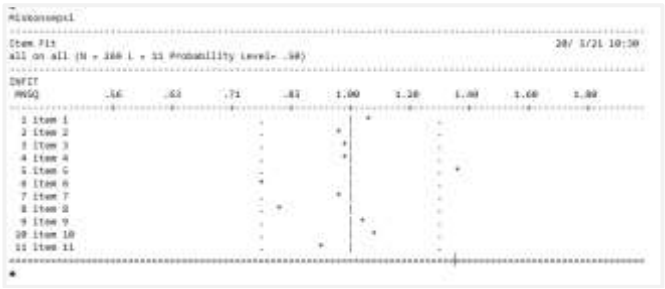

Figure 5. Distribution of INFIT MNSQ values for diagnostic test items

Figure 5 shows the distribution of INFIT MNSQ values for Items 1 to 11 in the diagnostic test instrument. According to standard interpretation criteria (Bond, 2015) acceptable INFIT MNSQ values lie within the range of 0.77 to 1.30. Based on this range, 10 out of the 11 items in the test are within acceptable limits. One exception was Item 5, which recorded an INFIT MNSQ of 1.35, slightly above the upper threshold. However, because the deviation is minimal (only 0.05) and not significantly beyond the model's tolerance range, the item was retained with minor revisions. Its inclusion is still considered acceptable for further development and analysis.

Reliability

Reliability is a fundamental criterion for assessing whether a test instrument is feasible and appropriate for use in measurement contexts. An instrument is considered reliable if it consistently yields stable and reproducible results across repeated administrations or samples (Suantari et al., 2018). In this study, the reliability of the diagnostic misconception test was analyzed using the Quest software, which provides an estimate of the test's internal consistency based on Item Response Theory (IRT) parameters. The obtained reliability coefficient was 0.71, which falls into the moderate reliability category. This finding is consistent with the interpretation provided by George et al. (2020), who suggested that reliability coefficients in the range of 0.70 to 0.79 indicate acceptable reliability for research instruments. A reliability coefficient of 0.71 suggests that the instrument is adequately stable, with a reasonable degree of consistency in students' responses across items.

Furthermore, a higher reliability value generally indicates that more items and more participants contribute to the precision of the measurement. Conversely, lower reliability suggests that fewer test takers or items are effectively contributing to measurement accuracy (Prihatni et al., 2016). Given this result, the diagnostic misconception test can be considered sufficiently reliable for the purpose of identifying students' misconceptions in the topic of sound and light waves. However, future refinements and broader field trials could further enhance the instrument's measurement precision.

Item Difficulty Level

Item difficulty analysis was conducted to evaluate whether the test items were appropriately challenging for the target group and capable of differentiating students based on their conceptual understanding. This is essential to ensure that the diagnostic test effectively reveals students' misconceptions, without being overly easy or difficult (Suantari et al., 2018).

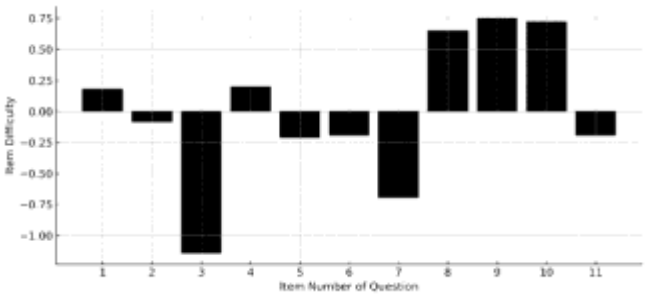

Figure 6. Difficulty levels for the diagnostic misconception

Test Figure 6 illustrates the distribution of item difficulty levels derived from the empirical trial. Based on the analysis, item 9 was identified as the most difficult, with a difficulty parameter of 0.75, whereas item 3 was the easiest, with a difficulty level of –1.14. According to Item Response Theory (IRT), particularly within the Partial Credit Model (PCM), a test item is considered to have an acceptable difficulty level if its difficulty parameter falls within the range of –2 to +2 (Istiyono, 2018).

In general, for diagnostic purposes, effective test items are those that are moderately difficult, meaning they are neither too easy nor too hard (Tumanggor et al., 2020). This balance allows the test to better differentiate students' understanding and uncover misconceptions. The observed range of item difficulty in this study confirms that the test items are well-targeted for the student population.

Item Characteristic Curve

Item characteristics can be visualized through an item characteristic curve (ICC), which provides insight into the probability of students with varying ability levels responding to each item category. This curve was generated using the Parscale software, which is suitable for analyzing polytomous items under the Partial Credit Model (PCM).

The x-axis of the curve represents students' latent ability levels, ranging from -3 to +3, while the y-axis indicates the probability of students obtaining a score in a given category (ranging from 0 to 4) for a specific item.

The ICC allows researchers to evaluate how well each item discriminates between students of different abilities and how the scoring categories function across the ability continuum.

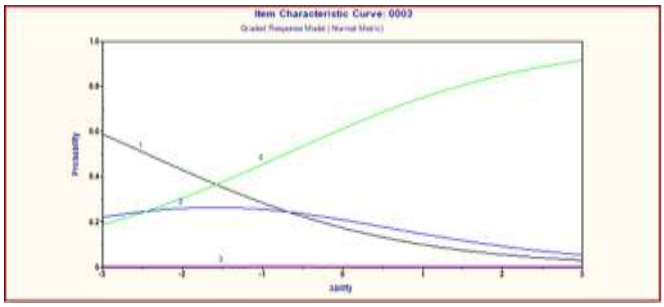

Figure 7. Item characteristic curve for item number 3

Figure 7 presents the item characteristic curve for item number 3. The curve shows that students with low ability have the highest probability of obtaining a score of 1. As student ability increases, the probability of achieving lower scores (such as 1 and 2) decreases, while the probability of attaining a higher score (such as 4) increases. The curves for intermediate scores (such as 2 and 3) indicate transitional functions, representing students with moderate conceptual understanding. This pattern suggests that the item effectively discriminates between students at different levels of ability, which aligns with the principles of polytomous item response theory (IRT).

Information Function dan Standard Error of Measurement

One key strength of models such as the Partial Credit Model (PCM) is their ability to estimate how much information a test provides across varying levels of student ability. This is represented by the information function curve, which helps determine the precision of the instrument at different ability levels. The Standard Error of Measurement (SEM), in contrast, indicates the degree of uncertainty or potential error in those ability estimates. By analyzing both curves, developers can evaluate where the test is most and least effective.

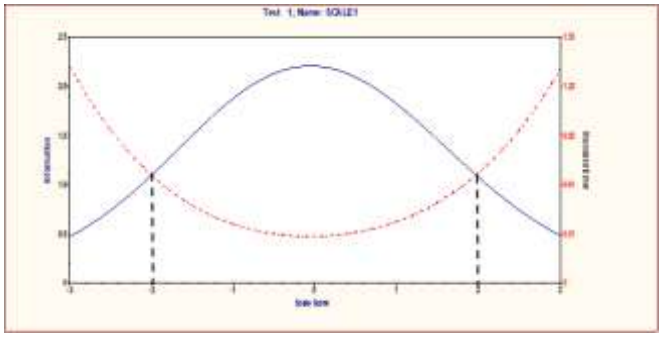

Figure 8. Relationship between information function and standard error of measurement (SEM) in the diagnostic misconception test

Figure 8 illustrates the relationship between the information function and the Standard Error of Measurement, which is inversely proportional. As the information function increases, the SEM decreases, and vice versa. The intersection point of the information curve and the SEM curve represents the range within which the test provides the most reliable estimates of student ability (Andayani et al., 2019). From Figure 8, it can be seen that the diagnostic misconception test provides optimal information for students whose ability levels fall within the range of $-2 \leq \theta \leq 2$. This indicates that the instrument is most effective in measuring students within this moderate ability range and less effective for those with extremely low or high ability levels.

A limited trial was conducted to identify and revise or eliminate test items that did not fit the model. The analysis revealed that item number 5 showed misfit. To diagnose the source of this issue, further analysis was performed using the *tn* output from the QUEST program. After appropriate revisions, the item was re-tested and subsequently integrated into the PhyTestApp platform for implementation.

The psychometric findings established the robustness of the instrument, which subsequently informed the pilot implementation with students. This stage not only provided evidence of practicality but also offered insights from students' perspectives on the usability and relevance of the application.

Student and Teacher Responses to the Phytestapp Media

Figure 9 illustrates student responses to the PhyTestApp media across three assessment aspects: content and objectives, instructional effectiveness, and technical usability. Overall, the responses indicate a positive perception of the application. Students rated the content and objectives aspect the highest, suggesting that the material and learning goals are appropriate and aligned with the expected curriculum. Technical usability was also highly appreciated, showing that students found the application user-friendly and reliable. Instructional effectiveness received slightly lower ratings compared to the other aspects, indicating that while the application supports learning, there is still room for improvement in enhancing conceptual understanding and engagement. In summary, the data demonstrate that students generally responded favorably to the PhyTestApp media, highlighting its potential as a tool for facilitating learning and identifying misconceptions in science education.

Several student comments and suggestions were recorded, highlighting both strengths and areas for improvement. Many students stated that *PhyTestApp* is easy and practical to use. They emphasized that the application is very good, simple to operate, and practical

because it provides features ranging from learning materials to assessment in one platform. Even in its beta version, the app was considered impressive, requiring only minor refinements. Students also mentioned that the media was interesting, easy to understand, and helpful in making learning more accessible.

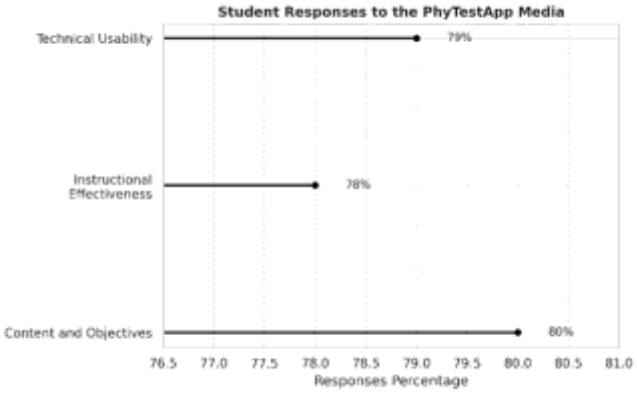

Figure 9. Student responses to the PhyTestApp media

Another strength noted was the availability of explanations after answering questions, which students found very useful. In addition, PhyTestApp was considered effective in aiding understanding of learning materials. The combination of educational videos and reading materials helped them grasp concepts more easily, and the variety of formats prevented boredom that often occurs with text-based assessments. Students also highlighted that the app supported their efforts in solving physics problems.

From a technical perspective, students appreciated that PhyTestApp required relatively small storage space, making it convenient to install and use. However, several weaknesses were also reported. Some users experienced errors, particularly when accidentally exiting the app, which caused it to get stuck on loading. Others noted issues when starting the test, emphasizing the need for improvements to prevent such disruptions in the future.

Finally, students provided feedback on the app's design. While acknowledging its usefulness in understanding the material, they felt that the visual design was not very appealing. Suggestions were made to make the app more interactive and engaging in order to enhance both its effectiveness and innovation.

In addition to student feedback, the PhyTestApp was also evaluated by physics teachers from the schools where the field tests were conducted. The response data were collected using a questionnaire that assessed the same aspects as those used in the student questionnaire, namely content and objectives, instructional effectiveness, and technical usability.

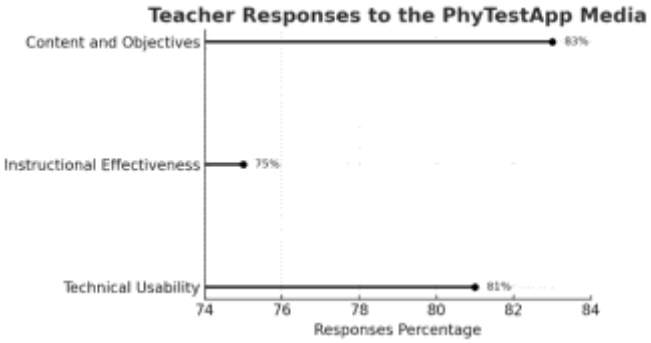

Figure 10. Teacher responses to the PhyTestApp media

Teacher responses to the PhyTestApp media are summarized in Figure 10. Overall, teachers expressed strong approval, particularly emphasizing the relevance of the content and the clarity of the learning objectives, which they viewed as well aligned with curriculum standards. In addition, the application's technical usability was highlighted as a strength, as teachers found the platform straightforward to operate and suitable for integration into classroom practice. While instructional effectiveness was rated somewhat lower than the other aspects, teachers still acknowledged its value in supporting lessons and recognized its potential for further development in fostering deeper student understanding and engagement. These insights suggest that, from a teacher's perspective, PhyTestApp not only addresses curriculum demands but also offers practical advantages for classroom implementation, with opportunities for refinement in its instructional features.

Teachers also provided several comments and suggestions regarding the PhyTestApp. They noted that the application is not yet compatible with iOS devices, which limits accessibility for some users. At the same time, they highlighted that the content and diagnostic feedback are highly useful and relevant for classroom implementation, particularly in identifying student misconceptions. The integration of video materials and question explanations was also appreciated, as these features support students' independent learning. In addition, teachers suggested improvements to the user interface design to make the application more appealing and intuitive for both teachers and students. Finally, they recommended the inclusion of analytics features in future versions to enable teachers to track student progress more effectively.

Media Revision

After the media validation stage, several improvements were made to enhance the overall quality and functionality of the PhyTestApp. Revisions focused on the user interface, technical performance, and usability based on feedback from both students and teachers. Suggestions included enhancing the visual

design to make it more appealing, providing a clearer explanation on the homepage regarding the purpose and use of the application, and automating the registration process to simplify user access.

In addition, technical issues reported by users, such as application freezing when accidentally exited or loading delays during test initiation, were addressed. These improvements aimed to increase the reliability, practicality, and user satisfaction of the PhyTestApp. After completing the revisions, the updated media was finalized and deemed ready for implementation as a digital diagnostic tool to help identify students' misconceptions in the topics of sound and light waves.

Final Product Dissemination Stage

The final version of the PhyTestApp media was disseminated to physics teachers who are members of the MGMP (Subject Teacher Consultation Forum) community in Yogyakarta. This distribution aimed to expand the use of the application as a digital tool to identify and address students' misconceptions in physics learning. Furthermore, the results of this research and development have been compiled into a scholarly article that will be submitted for publication in an educational journal, with the aim of contributing to academic discourse and promoting the adoption of digital diagnostic tools in science education.

Conclusion

This study successfully developed the Physics Test Application (PhyTestApp) as a diagnostic tool to identify students' misconceptions in science. The instrument, consisting of two-tier multiple-choice items, demonstrated strong content validity and acceptable reliability, confirming its effectiveness in measuring various levels of student understanding. Usability testing also indicated that the application is practical, feasible, and positively received by both students and teachers. Overall, PhyTestApp provides a reliable and innovative approach to diagnostic assessment by integrating psychometric modeling with mobile technology. Its ability to deliver immediate feedback and support targeted instruction highlights its potential as a valuable tool for improving science education and addressing misconceptions in line with 21st-century learning goals.

Acknowledgments
Thank you to all parties who have helped in this research so that this article can be published.

Author Contributions
All authors contributed to writing this article.

Funding
No external funding.

Conflicts of Interest
No conflict interest.

References

Abidin, Z. A. (2018). *Pengembangan Computerized Adaptive Test (CAT) untuk Memetakan Keterampilan Berpikir Kritis Fisika Peserta Didik Kelas XI SMA*. Universitas Negeri Yogyakarta.

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

Akbar. (2013). *Instrumen Perangkat Pembelajaran*. PT. Remaja Rosdakarya.

Andayani, A., & Ramalis, T. R. (2019). Kajian implementasi teori respon butir dalam menganalisis instrumen tes materi fisika. *Prosiding Seminar Nasional Fisika 5.0*, 37–42. Retrieved from https://fisika.upi.edu/wp-content/uploads/2020/02/Prosiding-Sinafi-2019.pdf

Çelikkanlı, N. Ö., & Kızılcık, H. Ş. (2022). A Review of Studies About Four-Tier Diagnostic Tests in Physics Education. *Journal of Turkish Science Education*, *19*(4), 1291–1311. https://doi.org/10.36681/tused.2022.175

Chen, C. H. H. I. H., Lin, H. U. H., & Lin, M. I. N. G. I. (2003). Developing a Two-Tier Diagnostic Instrument to Assess High School Students ' Understanding − The Formation of Images by a Plane Mirror. *Proceedings of the National Science Council*, *12*(3), 106–121. Retrieved from https://shorturl.at/AiRkc

Dellantonio, S., & Pastore, L. (2021). Ignorance, misconceptions and critical thinking. *Synthese*, *198*(8), 7473–7501. https://doi.org/10.1007/s11229-019-02529-7

George, D., & Mallery, P. (2020). *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference (Sixteenth)*. Routledge. https://doi.org/10.4324/9780429056765

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science.

Eurasia Journal of Mathematics, Science and Technology Education, *11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Istiyono, E. (2018). *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika*. UNY Press.

Juita, Z., Sundari, P. D., Sari, S. Y., & Rahim, F. R. (2023). Identification of Physics Misconceptions Using Five-tier Diagnostic Test: Newton's Law of Gravitation Context. *Jurnal Penelitian Pendidikan IPA*, *9*(8), 5954–5963. https://doi.org/10.29303/jppipa.v9i8.3147

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science and Technological Education*, *35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094

Laliyo, L. A. R., Botutihe, D. N., & Panigoro, C. (2019). The development of two-tier instrument based on distractor to assess conceptual understanding level and student misconceptions in explaining redox reactions. *International Journal of Learning, Teaching and Educational Research*, *18*(9), 216–237. https://doi.org/10.26803/ijlter.18.9.12

Nainggolan, J., Silaban, B., Sinaga, D., & Zendrato, F. (2023). Analysis of Physics Misconceptions of Students in Mechanic Materials Using the Tier Multiple Choice Diagnostic Test. *AL-ISHLAH: Jurnal Pendidikan*, *15*(3), 3578–3586. https://doi.org/10.35445/alishlah.v15i3.3023

Nugraeni, D., Jamzuri, J., & Sarwanto, S. (2013). Penyusunan tes diagnostik fisika materi listrik dinamis. *Jurnal Pendidikan Fisika*, *1*(2), 12–16. Retrieved from http://jurnal.fkip.uns.ac.id/index.php/pfisika/article/view/2796

Nurhasanah, Hidayatullah, Z., Badrus, M., & Arif, S. (2024). Karakteristik Instrumen Tes Literasi Digital Ditinjau dari Validitas Isi dan Validitas Empiris (Kecocokan Butir dengan Model, Reliabilitas, serta Tingkat Kesukaran Butir). *Journal of Classroom Action Research*, *6*(4), 917–923. https://doi.org/10.29303/jcar.v6i4.9650

Oriondo, & Dallo-Antonio. (1984). *Evaluating Educational outcomes (test, measurement, and evalution)*. REX Printing Company, Inc.

Prihatni, Y., Kumaidi, K., & Mundilarto, M. (2016). Pengembangan instrumen diagnostik kognitif pada mata pelajaran IPA di SMP. *Jurnal Penelitian dan Evaluasi Pendidikan, 20*(1), 111-125. Retrieved from http://journal.uny.ac.id/index.php/jpep. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(1), 111–125.

Pujayanto, P., Budiharti, R., Adhitama, E., Nuraini, N. R. A., & Putri, H. V. (2018). The development of a web-based assessment system to identify students' misconception automatically on linear kinematics with a four-tier instrument test. *Physics Education*, *53*(4). https://doi.org/10.1088/1361-6552/aac695

Rohmah, Z., Handhika, J., & Huriawati, F. (2018). E-diagnostic test untuk mengungkap miskonsepsi kinematika. *SPEKTRA : Jurnal Kajian Pendidikan Sains*, *4*(2), 162. https://doi.org/10.32699/spektra.v4i2.57

Sahidu, H., Gunawan, G., Indriaturrahmi, I., & Astutik, F. (2017). Desain Sistem E-Assessment Pada Pembelajaran Fisika Di Lptk. *Jurnal Pendidikan Fisika Dan Teknologi*, *3*(2), 265. https://doi.org/10.29303/jpft.v3i2.422

Salma, V. M. (2015). *Pengembangan E-Diagnostic Test Untuk Mengidentifikasi Pemahaman Konsep Fisika Siswa Sma Pada Pokok Bahasan Fluida Statis* Universitas Negeri Semarang. https://doi.org/10.15294/upej.v5i1.12701

Samsudin, A., Zulfikar, A., Saepuzaman, D., Suhandi, A., Aminudin, A. H., Supriyadi, S., & Coştu, B. (2024). Correcting grade 11 students' misconceptions of the concept of force through the conceptual change model (CCM) with PDEODE*E tasks. *Journal of Turkish Science Education*, *21*(2), 212–231. https://doi.org/10.36681/tused.2024.012

Sarasvati, A. (2016). *Pengembangan Science Assesment Website (Sc-Wb) Tema Sistem Ekskresi Manusia untuk Kelas VIII SMP*. Universitas Negeri Semarang. Retrieved from http://lib.unnes.ac.id/28833/

Soehato, S., & Csapo, B. (2021). Evaluating Item Difficulty Patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. *Heliyon*, *7*. https://doi.org/10.1016/j.heliyon.2021.e08352

Suantari, N. M. D. P., Suma, K., & Pujani, N. M. (2018). Development of three tier static electricity diagnostic test to identify student conceptions about static electricity. *International Research Journal of Engineering, IT & Scientific Research*, *4*(5). https://doi.org/10.21744/irjeis.v4n5.285

Suban, M. E., Hidayatullah, Z., & Nurhasanah. (2024). Identifikasi Miskonsepsi Menggunakan Three-Tier Diagnostic Test dan Representasi Gambar pada Konsep Gaya. *Hamzanwadi Journal of Science Education,* *1*(2), 1–9. https://doi.org/10.29408/hijase.v1i2.26917

Sukatiman, Saputro, I. N., & Budiarto, M. K. (2024). Digital Classroom Innovations: Leveraging Smartphone-Based Application To Stimulate Students Creative Thinking Skills. *Journal on Efficiency and Responsibility in Education and Science*, *17*(4), 349–360. https://doi.org/10.7160/eriesj.2024.170407

Suparman, A. R., Rohaeti, E., & Wening, S. (2024). Development of Computer-Based Chemical Five-Tier Diagnostic Test Instruments: a Generalized Partial Credit Model. *Journal on Efficiency and Responsibility in Education and Science*, *17*(1), 92–106. https://doi.org/10.7160/eriesj.2024.170108

Susanti, M., Rusilowati, A., & Susanto, H. (2015). Pengembangan Bahan Ajar Ipa Berbasis Literasi Sains Bertema Listrik Dalam Kehidupan Untuk Kelas IX. *Unnes Physics Education Journal*, *4*(3). https://doi.org/10.15294/upej.v4i3.9973

Syamsuddin, S. (2023). Implementasi Classic Test dan Item Respon Theory Pada Penilaian Tes Pembelajaran Matematika. *EDUSCOPE: Jurnal Pendidikan, Pembelajaran, Dan Teknologi*, *8*(2), 28–43. https://doi.org/10.32764/eduscope.v8i2.3488

Taslidere, E. (2016). Development and Use of a Three-tier Diagnostic Test to Assess High School Students' Misconceptions about the Photoelectric Effect. *Research in Science and Technological Education*, *34*(2), 164–186. https://doi.org/10.1080/02635143.2015.1124409

Theobald, M., & Brod, G. (2021). Tackling Scientific Misconceptions: The Element of Surprise. *Child Development*, *92*(5), 2128–2141. https://doi.org/10.1111/cdev.13582

Thiagarajan, S. (1974). *Instructional Development for Training Teachers of Exceptional Children*. Sourcebook.

Tsui, C. Y., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, *32*(8), 1073–1098. https://doi.org/10.1080/09500690902951429

Tumanggor, A. M. R., Supahar, S., Ringo, E. S., & Harliadi, M. D. (2020). Detecting Students' Misconception in Simple Harmonic Motion Concepts Using Four-Tier Diagnostic Test Instruments. *Jurnal Ilmiah Pendidikan Fisika Al-Biruni*, *9*(1), 21–31. https://doi.org/10.24042/jipfalbiruni.v9i1.4571

Vosniadou, S. (2020). *Students' Misconceptions and Science Education*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.965

Wahidah, N., & Saptono, S. (2019). The Development of Three Tier Multiple Choice Test to Explore Junior High School Students' Scientific Literacy Misconceptions. *Journal of Innovative Science Education, 8*(2), 190-198. Retrieved from https://journal.unnes.ac.id/sju/index.php/jise/article/view/27927

Wahyuningsih, T., Raharjo, T., & Fitriana Masithoh, D. (2013). Pembuatan Instrumen Tes Diagnostik Fisika SMA Kelas XI. *Jurnal Pendidikan Fisika*, *1*(1), 111–117. Retrieved from https://jurnal.fkip.uns.ac.id/index.php/pfisika/article/view/1785