Jurnal Penelitian dan Evaluasi Pendidikan
Volume 29, No. 1, June 2025 (71-84)
Online: http://journal.uny.ac.id/index.php/jpep

Psychometric evaluation of diagnostic test instruments on physics wave material

Muh. Asriadi AM[1]*; Anna Isabela Sanam[2]
[1]Universitas Pendidikan Indonesia, Indonesia
[2]Institute of Business, Timor-Leste
*Corresponding Author. E-mail: muhasriadi@upi.edu

ARTICLE INFO

Article History
Submitted:
23 April 2025
Revised:
24 April
Accepted:
25 April

Keywords
aiken's v; efa; cfa; reliability; diagnostic test

Scan Me:

ABSTRACT

This study aims to evaluate the psychometric quality of diagnostic test instruments on wave material in physics learning, with a focus on their validity and reliability. The instrument was validated by seven experts using the content validation method, which resulted in an average validity score of 0.849, indicating that the instrument is very valid. Confirmatory factor analysis (CFA) was conducted to test the model fit and validity of the instrument structure. The KMO test results showed a figure of 0.946, indicating the suitability of the data for factor analysis, while the Bartlett Test of Sphericity results confirmed that the intercorrelation data was not an identity matrix. The tested model showed a good fit to the data, with a CFI value of 0.948, TLI of 0.935, and RMSEA of 0.072. The model parameter estimates showed a significant relationship between the latent variables and the observed items, with a strong influence from the "Wave" variable. Although the Chi-square test results indicated a statistically poor fit of the model, this result could be influenced by the large sample size. Limitations of this study include the limited sample size of seven validators, which may not cover all perspectives. This study suggests that further research involving a larger number of validators and more in-depth analysis should be conducted to strengthen the results obtained. The results of this study are expected to contribute to the development of more effective diagnostic test instruments for physics teaching at the secondary school level.

To cite this article (in APA style):
AM, M. A., & Sanam, A. I. (2025). Psychometric evaluation of diagnostic test instruments on physics wave material. *Jurnal Penelitian dan Evaluasi Pendidikan, 29(1)*, 71-84 doi: https://doi.org/10.21831/pep.v29i1.84655

INTRODUCTION

Diagnostic test instruments serve as essential tools in physics education, particularly for identifying misconceptions and assessing students' conceptual understanding more deeply (Hyland & O'Shea, 2022; Tomczyk & Eger, 2020). Unlike summative tests that only assess final learning outcomes, diagnostic tests serve as a teacher's tool to map students' initial abilities and specifically identify parts of the concept that have not been understood (Homjan et al., 2022). In the context of physics learning, this is very important because physics requires not only memorization of concepts but also logical understanding, relationships between concepts, and scientific reasoning skills (Hadi et al., 2022). One of the materials that is known to be complex and often cause misconceptions is wave material, which involves abstract concepts such as amplitude, frequency, wavelength, phase and interference. Therefore, a diagnostic instrument is needed that is not only accurate in its content but also has a strong scientific basis in terms of validity and reliability.

Unfortunately, conditions in the field often show that the instruments used in learning are still far from ideal. Several previous studies have shown that diagnostic test instruments circulating in schools tend to be made practically without going through an adequate validation process. For example, research by Istiyono (2022) and Burkholder et al. (2021) found that most teachers only used homemade questions without empirical trials, potentially producing

10.21831/pep.v29i1.84655

inaccurate misconception data. In addition, a study by Scoulas et al. (2021) showed that the construct validity of many diagnostic instruments has not been systematically analyzed through an exploratory or confirmatory factor approach. In fact, validity and reliability are absolute requirements for an instrument to be said to be appropriate and valid for use in the context of educational assessment (Asriadi & Hadi, 2021; Tabatabaee-Yazdi et al., 2018).

This gap indicates the need for the development and evaluation of diagnostic instruments based on modern psychometric approaches. The developed instruments must go through a content validation process using quantitative approaches such as Aiken's V index, followed by exploratory factor analysis (EFA) to identify the latent structure of the construct (Burak & Gültekin, 2021) and confirmatory factor analysis (CFA) to test the suitability of the model to empirical data (Abdullah et al., 2021). Reliability estimation is also important to ensure the internal consistency of the test items (Haryanto et al., 2023). This kind of psychometric evaluation has not been systematically conducted in the context of wave material diagnostic instruments, thus becoming an important opportunity to produce significant scientific contributions.

This study aims to evaluate the psychometric quality of diagnostic test instruments on wave material in physics learning, with a focus on their validity and reliability. In ensuring a clear and coherent research direction, this study focuses on two main objectives: (1) assessing the validity of the content analyzed using the Aiken's V index to determine the alignment of test items with learning indicators and content clarity while constructing validity is examined through Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) to empirically validate the dimensional structure of students' conceptual understanding of waves, and (2) evaluating the internal consistency of the instrument to ensure its reliability in identifying misconceptions. This study builds on existing research in physics education that highlights the need for valid and reliable diagnostic tools to support formative assessment and guide targeted instructional interventions. Through this approach, this study is expected to contribute to a state-of-the-art diagnostic instrument that supports evidence-based decision-making in the learning process.

RESEARCH METHOD

This study uses a survey design with a cross-sectional approach, in which data is collected at a single point in time (Creswell, 2022). This approach was chosen because it, in accordance with the main objective of the study, conducts a psychometric evaluation of diagnostic test instruments on wave material in physics learning. The survey design allows researchers to obtain quantitative data from large numbers of respondents efficiently, which is important to support the analysis of the validity and reliability of the instrument (Johnson & Christensen, 2017). The cross-sectional approach is very relevant in the context of measuring psychometric properties because it allows researchers to test the construction structure of the instrument (through Exploratory Factor Analysis and Confirmatory Factor Analysis), as well as estimate reliability (with internal consistency coefficients) based on actual data from the target population in a short time. Thus, this method provides a strong empirical picture of the extent to which the instrument can measure diagnostic concepts validly and reliably according to the characteristics of students at the time the measurement is carried out.

The subjects in this study were 11th-grade students from various SMA/MAN in Bandung City, West Java. This study involved two stages of data collection. In the initial stage, a total of 220 students were planned to be tested using the diagnostic test instrument. The sample at this stage was selected using a simple random sampling technique to ensure the representativeness and diversity of respondents from the student population in Bandung City. This technique allows each student to have an equal opportunity to be selected as a subject so that the data

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

obtained can describe a more general condition. The selected sample is expected to provide a representative picture of the effectiveness of the diagnostic test instrument being tested.

The instrument used in this study was a diagnostic test designed to identify and evaluate the level of initial understanding and development of student's abilities in the material being taught. This test uses a two-tier multiple-choice format that aims to measure students' understanding at various cognitive levels based on Bloom's Taxonomy. The material tested includes Sound Waves and Light Waves, with questions categorized into three cognitive levels: C2 (Understanding), C3 (Applying), and C4 (Analyzing).

The matrix of diagnostic test instruments used can be seen in Table 1. This table shows the distribution of questions based on the material and cognitive level that is in accordance with the measurement objectives to be achieved. By using this approach, the test aims to measure students' understanding at various cognitive levels, from basic knowledge to the level of analysis and evaluation. The total number of questions in this test is 14 questions designed to represent the various cognitive levels in Bloom's Taxonomy.

Table 1. Diagnostic Test Instrument Matrix (Two-Tier Multiple Choice)

Cognitive Levels	Material							
	Vibrations & Waves		Sound Wave 1		Sound Wave 2		Light Waves	
	Question Number	Number of Questions	Question Number	Number of Questions	Question Number	Number of Questions	Question Number	Number of Questions
C2	3	1	8	1	9	1	14	1
C3	2	1	5	1	10	1	12	1
C4	1.4	2	6.7	2	11	1	13	1
Total	4		4		3		3	

Table 2. Results of Validation of Final Diagnostic Test Instrument Content (Post-test)

Item	R1	R2	R3	R4	R5	R6	R7	V	Description
Item 1	5	4	5	5	4	4	3	0.821	Valid
Item 2	5	5	3	4	5	5	3	0.821	Valid
Item 3	5	3	5	5	5	5	3	0.857	Valid
Item 4	5	5	5	4	5	5	4	0.928	Valid
Item 5	5	4	4	3	5	4	4	0.786	Valid
Article 6	4	5	5	4	5	4	3	0.821	Valid
Item 7	5	4	5	3	4	5	4	0.821	Valid
Article 8	5	5	5	4	5	5	4	0.928	Valid
Article 9	5	5	5	4	4	4	4	0.786	Valid
Article 10	4	4	5	4	5	5	4	0.857	Valid
Article 11	5	4	4	4	5	4	5	0.893	Valid
Article 12	4	5	3	4	5	5	4	0.821	Valid
Article 13	5	4	4	5	4	5	5	0.928	Valid
Article 14	4	5	5	3	5	5	3	0.821	Valid
Validation Results of the Contents of the Diagnostic Test Instrument								0.849	Valid

With this approach, the test instrument not only aims to measure students' understanding at the basic knowledge level but also to assess students' ability to apply concepts and analyze more complex physics phenomena according to the expected cognitive level. Based on information table 1 is data analysis in this study involved three main stages: content validity, construct validity, and instrument reliability. First, content validity was tested using assessments from expert validators and practitioners through instrument validation sheets. The scores given ranged from 1 to 5, with categories ranging from irrelevant to very relevant, and then analyzed using the Aiken formula to measure the level of validity. The results were categorized as very valid, valid, or less valid according to the Aiken coefficient. Furthermore, construct validity was

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

tested using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) Order 2 to ensure that the selected test items had a factor load above 0.5 to be retained as valid. Items with factor loads between 0.30 and 0.50 would be considered for further analysis (Hair et al., 2017). Finally, the reliability of the instrument was assessed using McDonald's ω and Cronbach's α, which are appropriate methods for estimating the reliability of instruments with polytomous scale items (Blessing et al., 2021). The results are categorized into five levels of reliability based on Cronbach's Alpha values, ranging from unreliable to highly reliable. All these analyses were performed using R Studio software.

FINDINGS AND DISCUSSION

Findings

Content Validation Results

The results of the content validation of the diagnostic test instrument show that the assessment data by 7 expert validators and practitioners, obtained through the instrument validation sheet, were analyzed to assess the content validity of the developed diagnostic test. This content validity focuses on the suitability of the question items to the established indicators. The assessment was carried out by measurement experts and material experts, where each question item was given a score between 1 and 5, with categories of irrelevant, less relevant, relevant, and very relevant. The scores obtained from each validator were then calculated using the Aiken formula to determine the content validity of the instrument.

Table 2 shows the results of the validation of the final diagnostic test instrument based on expert assessments. Each item was tested and had a validity score (V) varying between 0.786 and 0.928, all of which were in the Valid category. Overall, the average validity value of this instrument reached 0.849, which was also included in the Valid category. Based on these results, it can be concluded that this final diagnostic test instrument has met the criteria for very good validity and can be used with confidence to measure the desired aspects.

KMO and Bartlett test of Sphericity

The KMO (Kaiser-Meyer-Olkin) test and the Bartlett test (test of sphericity) are two important tests in preparing for factor analysis. The KMO test is used to assess whether the data is suitable for factor analysis by looking at the correlation between variables, while the Bartlett test tests the significance of the correlation between variables (Watkins, 2018) Appropriate results from both tests are needed before proceeding to further factor analysis.

Table 3. KMO and Bartlett Test of Sphericity

Parameter		Mark
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.946
Bartlett's Test of Sphericity	Approx. Chi-square	1639.326
	df	91
	Sig.	.000

The results in Table 3 show that the KMO value has met the requirements, so the data is worthy of further analysis. The eligibility criteria are that the KMO value must be more than 0.7, or at least 0.5, with a Bartlett significance value below 0.05. The Bartlett tests whether the intercorrelation matrix is an identity matrix or not. If the significance value is <0.05, then the

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

intercorrelation matrix is not an identity matrix, and factor analysis can be performed. The results of the analysis show that the KMO value is 0.946, which is greater than 0.7, and the significance value is 0.000, which is less than 0.05. Thus, the intercorrelation matrix is not an identity matrix, allowing factor analysis to be performed.

MSA (Measure of Sampling Adequacy)

Measure of Sampling Adequacy (MSA) is a measure used in factor analysis to evaluate the extent to which sample data is suitable or adequate for factor analysis. MSA calculates the partial correlation between variables in a dataset, which is then used to determine whether the variables can be grouped or reduced into smaller factors. MSA values range from 0 to 1, where values close to 1 indicate that the variables in the dataset have a high correlation and are very suitable for factor analysis, while values close to 0 indicate a low correlation and data that is less suitable for factor analysis. Interpretation of MSA values is important to ensure the validity and reliability of factor analysis that will be carried out on existing sample data.

Table 4. MSA (Measure of Sampling Adequacy)

Item Parameters	MSA Value	Item Parameters	MSA Value
Item_1	0.926	Item_9	0.959
Item_2	0.912	Item_10	0.964
Item_3	0.957	Item_11	0.927
Item_4	0.937	Item_12	0.958
Item_5	0.925	Item_13	0.963
Item_6	0.949	Item_14	0.959
Item_7	0.957	Overall MSA	0.946
Item_8	0.952		

Based on the analysis results in table 4, all items have MSA values above 0.5. Items with fairly high MSA values include Item 1 with a value of 0.926, Item 2 with a value of 0.912, and so on up to Item 14 which has a value of 0.959. Therefore, all items meet the sample adequacy requirements for further analysis.

Model Fit Test

Model Fit Test is an evaluation process carried out in data analysis, especially in the context of statistical models such as SEM (Structural Equation Modeling) (Helmi et al., 2025). This test aims to assess the extent to which the proposed model is in accordance with the observed data.

Based on the results of the model analysis presented in Table 5, it can be concluded that this model shows a good level of fit to the data analyzed. Although the Chi-square test shows a low p-value (0), indicating that the model does not fit perfectly based on this criterion, this needs to be considered carefully because a large sample size can affect the results of this test (Hair et al., 2019) . However, several other fit indices show very positive results. The Comparative Fit Index (CFI) of 0.948 and the Tucker-Lewis Index (TLI) of 0.935 both exceed the recommended cut-off ($\geq$ 0.90), indicating that this model is very suitable for the data used. In addition, the Root Mean Square Error of Approximation (RMSEA) value of 0.072 and the Standardized Root Mean Square Residual (SRMR) of 0.043 are within adequate limits ($\leq$ 0.08), indicating that the model has a good statistical fit to the data. In addition, the Akaike Information Criterion (AIC) value of 8002.046 and the Bayesian Information Criterion (BIC) of 8110.497 indicate that this model is relatively better compared to other alternative models in terms of fit to the data (Hair

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

et al., 2017). Thus, although there are notes related to the Chi-square test, overall, this model can be considered a good representation of the observed phenomenon based on the various evaluation criteria used.

Table 5. Model Fit Test

Model Parameters	Cut Off Value	Mark	Description
Statistical test	-	155.926	Fit
Degrees of freedom	-	73	
P-value (Chi-square)	> 0.05	0	Fit
Comparative Fit Index (CFI)	≥ 0.90	0.948	Fit
Tucker-Lewis Index (TLI)	≥ 0.90	0.935	Fit
Akaike (AIC)	Smaller is better	8002.046	Fit
Bayesian (BIC)	Smaller is better	8110.497	Fit
Root Mean Square Error of Approximation (RMSEA)	≤ 0.08	0.072	Fit
Standardized Root Mean Square Residual (SRMR)	≤ 0.08	0.043	Fit

Compared to previous studies examining diagnostic instruments in science education, such as that conducted by Watkins (2018), this study offers an integrated approach that not only strengthens the internal structure of the instrument but also provides empirical support for its effectiveness in identifying students' misconceptions related to the concept of waves. The novelty of this study lies in its specific focus on wave phenomena in physics that are often underrepresented in the development of diagnostic tools. Furthermore, the use of rigorous statistical criteria ensures a strong validation of the instrument. This contribution lays the groundwork for future research aimed at developing similarly structured instruments in other complex domains in science education.

Model Parameter Estimation

Table 6 shows the results of estimation and statistical analysis for various latent variables in a model. Each row in the table lists estimates for each item, indicating how much influence that item has on the unobserved latent variables such as Item_1 to Item_14**.** Based on the result table 6 higher estimates indicate a stronger relationship between the item and the latent variable. In addition, there are also estimates for the effects of other latent variables, such as M1, M2, M3, and M4, on the observed latent variable. Positive values indicate a positive relationship, while negative values indicate a negative relationship between these latent variables. One significant result is the effect of the latent variable "Wave" on the observed latent variable, with an estimate of 0.232. This value is expressed by a Standard Error of 0.054 and a z-value of 4.341, indicating that the effect of "Wave" on the observed latent variable is statistically significant. This analysis is also supported by a very small P value, close to zero, indicating that this result is unlikely to occur by chance. The entire table provides a detailed picture of the relative contribution of each latent variable and item to the overall model, as well as the level of statistical significance of the given estimates.

R Square (Effect Size)

Table 7 shows the R Square (Effect Size) values for various latent variables in a statistical model. R Square indicates how much of the variation of a latent variable can be explained by

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

other variables in the model. Each row in the table lists the R Square estimates for each item, such as Item_1 to Item_14, as well as other latent variables such as M1, M3, and M4.

Table 6. Model Parameter Estimation

Latent Variables	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
Item_1	0.519	0.053	9.883	0	0.519	0.668
Item_2	0.605	0.075	8.091	0	0.605	0.367
Item_3	0.434	0.044	9.897	0	0.434	0.673
Item_4	0.726	0.082	8.801	0	0.726	0.439
Item_5	0.698	0.073	9.562	0	0.698	0.444
Item_6	0.501	0.057	8.749	0	0.501	0.312
Item_7	0.792	0.080	9.901	0	0.792	0.550
Item_8	0.468	0.054	8.703	0	0.468	0.308
Item_9	0.583	0.068	8.606	0	0.583	0.475
Item_10	0.608	0.070	8.743	0	0.608	0.490
Item_11	0.733	0.076	9.700	0	0.733	0.648
Item_12	0.549	0.073	7.486	0	0.549	0.360
Item_13	0.590	0.067	8,816	0	0.590	0.472
Item_14	1,022	0.102	10.051	0	1.022	0.747
M1	0.026	0.012	2.058	0.040	0.099	0.099
M2	-0.025	0.026	-0.975	0.330	-0.029	-0.029
M3	0.051	0.037	1.376	0.169	0.080	0.080
M4	0.088	0.053	1.677	0.093	0.090	0.090
Wave	0.232	0.054	4.341	0	1	1

Table 7. R Square (Effect Size)

Latent Variables	Estimate	Latent Variables	Estimate
Item_1	0.332	Item_10	0.510
Item_2	0.633	Item_11	0.352
Item_3	0.327	Item_12	0.640
Item_4	0.561	Item_13	0.528
Item_5	0.556	Item_14	0.253
Item_6	0.688	M1	0.901
Item_7	0.450	M2	NA
Item_8	0.692	M3	0.920
Item_9	0.525	M4	0.910

In table 7, several items, such as Item_6, Item_8, and Item_12, have quite high R Square values in succession. This indicates that these items make a large contribution in explaining the variation of the latent variables observed in this model. On the other hand, the latent variable M2 does not make a significant contribution or may not be included in the model for the observed latent variables, marked with the value "NA". This analysis provides an in-depth understanding of how well each latent variable and item can explain the variation of the observed latent variables, providing a strong foundation for interpreting the results and drawing conclusions in the context of the research or analysis being conducted.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

Path Standardized Value

Path standardized value is an adjusted value in structural equation modeling (SEM). This value indicates the strength and direction of the relationship between variables in the model after the variables are transformed into a standard form with a mean of zero and a standard deviation of one. The range of path standardized value values is from -1 to 1, where positive values indicate a positive relationship between the variables (i.e., when the independent variable increases, the dependent variable tends to increase as well), while negative values indicate a negative relationship (i.e., when the independent variable increases, the dependent variable tends to decrease). Values closer to 0 indicate a weak or insignificant relationship between the variables in the context of the analytical model being studied. Path standardized value is an important tool in understanding and interpreting how variables relate to each other in an analytical model, helping to test hypotheses and validate the developed model.

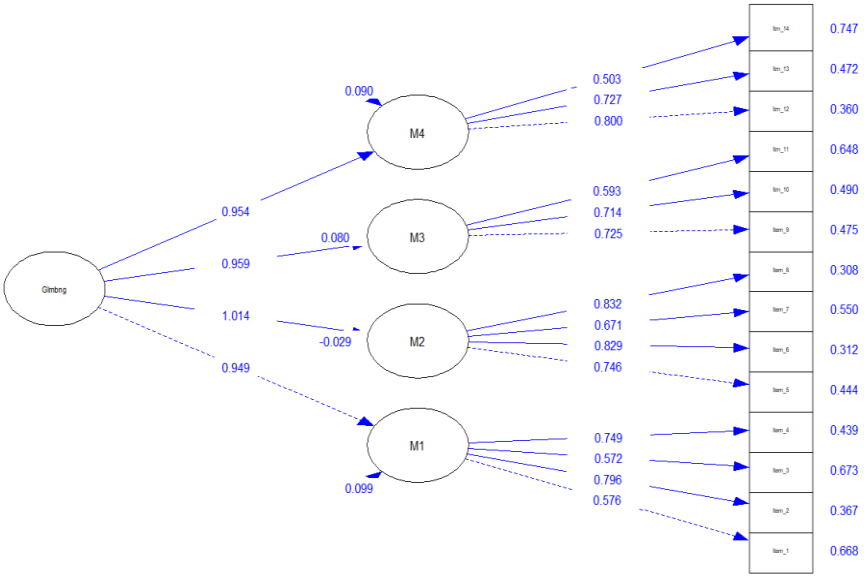

Figure 1. Path Standardized Value

The results of Figure 1 the confirmatory factor analysis (CFA) displayed in the Path Standardized Value diagram show a strong relationship between the main latent variable (wave) and the second latent variable (dimension). The loading factor from the wave variable to the dimension shows a value of more than 0.4, with a range between 0.949 to 1.014. This value indicates that the wave variable has a very strong relationship with its dimensions. In addition, the relationship from the dimensions to each question item also shows a significant loading factor, with a value of more than 0.4 ranging from 0.503 to 0.832. This shows that each dimension substantially explains the variance of the related question items. Overall, these results indicate that the model has a good fit, with the main and second latent variables explaining the variance of their indicators strongly and consistently.

Reliability Estimation Results

According to the reliability criteria, these values are included in the "Very High" category, because both exceed the threshold of 0.90. The 95% confidence interval for McDonald's Omega ranges from 0.917 to 0.944, while for Cronbach's Alpha it ranges from 0.912 to 0.939 (Navarro

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

et al., 2019). This shows that, with a 95% confidence level, the reliability of the instrument is estimated to be in a range that is still very high.

Table 8. Frequentist Scale Reliability Statistics

Estimate	McDonald's ω	Cronbach's α	Average interitem correlation
Point estimate	0.931	0.926	0.470
95% CI lower bound	0.917	0.912	0.425
95% CI upper bound	0.944	0.939	0.514

Based on information from table 8, the diagnostic test instrument shows a very high level of reliability. The point estimate value for McDonald's Omega (ω) is 0.931, and for Cronbach's Alpha (α) is 0.926. The average inter-time correlation is 0.470, with a 95% confidence interval ranging from 0.425 to 0.514. This high interitem correlation indicates that the items in the instrument are well correlated with each other, indicating that they measure the same construct with high consistency. Overall, these results indicate that the instrument has very good internal consistency, making it a reliable and consistent measuring instrument for the measurement purposes carried out.

Based on the statistical analysis of individual item reliability in Table 9, the diagnostic test instrument shows a very high level of internal consistency, according to the results of the McDonald's Omega (ω) and Cronbach's Alpha (α) tests.

Table 9. Frequentist Individual Item Reliability Statistics

Item	If item dropped	
	McDonald's ω	Cronbach's α
Item_1	0.929	0.925
Item_2	0.923	0.918
Item_3	0.929	0.925
Item_4	0.926	0.921
Item_5	0.924	0.919
Item_6	0.921	0.916
Item_7	0.927	0.922
Item_8	0.921	0.916
Item_9	0.926	0.921
Item_10	0.926	0.921
Item_11	0.929	0.925
Item_12	0.923	0.918
Item_13	0.926	0.921
Item_14	0.932	0.927

Based on the results table 9, the McDonald's ω values range from 0.921 to 0.932, while Cronbach's α values range from 0.916 to 0.927. This indicates that the instrument remains in the "Very High" reliability category even when one item is deleted, as all values remain above 0.90. Item-rest correlations, which measure the correlation of each item with the total score without that item, range from 0.477 to 0.804. The items with the highest correlations are Item 8 (0.804) and Item 6 (0.800), indicating that these two items are very consistent with the other items in measuring the same construct. Other items such as Item 2 (0.743) and Item 12 (0.747) also show very strong correlations with the total score, indicating significant contributions to the reliability of the instrument. On the other hand, Item 1 and Item 14 have lower item-rest correlations, at 0.524 and 0.477, respectively, but are still within the range indicating a positive contribution to the instrument's reliability. Overall, these results indicate that each item in the instrument contributes positively to the overall internal consistency. The high McDonald's ω

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

and Cronbach's α values despite the deletion of one item, as well as the generally strong item-rest correlations, indicate that this instrument is highly reliable for the purposes for which it was measured.

Discussion

An effective diagnostic test instrument in physics learning is very important in improving students' conceptual understanding. This instrument is not only used to measure the level of students' understanding of the material that has been studied but also functions to identify conceptual errors or misconceptions held by students. According to Shanmugam (2018), a good test instrument must have strong validity to measure what should be measured and high reliability so that the test results can be consistent if repeated. In the context of physics learning, the test instrument must cover all important concepts in physics, both conceptual questions and those involving practical applications of these concepts. Therefore, the development of diagnostic test instruments does not only rely on theory but also needs to be based on real observations of errors that often occur in students.

As part of the theory of instrument development, Martiana et al. (2022) stated that clear steps in the design are very important to create an effective test instrument. The stages consisting of planning, item testing, and evaluation aim to ensure that the instrument developed has high validity and reliability. Content validity needs to be considered when designing an instrument that focuses on the extent to which the questions given cover all the material that has been taught. Rahman et al. (2021) stated that to obtain good content validity, the development of questions must involve experts in the field, such as physicists or physics teachers. Thus, the questions designed not only measure students' understanding but also ensure that the scope of the material tested is by the applicable curriculum.

Previous research also shows that valid diagnostic test instruments can provide a clearer picture of students' misconceptions. Burkholder et al. (2021) found that the use of diagnostic tests designed with good validity can detect students' misconceptions more effectively, allowing teachers to respond immediately with the right approach. For example, by using diagnostic test results, teachers can find out which topics are difficult for students to understand and design more focused teaching on these aspects. Homjan et al. (2022) also revealed that test instruments based on content validity can increase evaluation effectiveness because the questions asked more accurately describe students' understanding of relevant physics concepts. Thus, a good test instrument not only functions as an evaluation tool but also as an instrument to facilitate continuous learning improvement.

In developing this diagnostic test instrument, one of the main challenges is to ensure that the questions developed do not only measure memory alone but also the ability of students to apply concepts in more complex situations (AM & Istiyono, 2022). In his educational taxonomy, Bloom explains that effective tests must cover various cognitive levels, from basic knowledge to application and analysis (Spiller & Tuten, 2019). Therefore, in developing a test instrument for physics, questions must be designed to measure not only the understanding of basic concepts but also the ability of students to solve more complex problems.

In addition, it is also important to consider the context of the use of this test instrument. Istiyono et al. (2023) show that in implementing diagnostic test instruments in the classroom, it is important to analyze the test results continuously. With this approach, teachers can reflect on the effectiveness of the learning that has been carried out and modify teaching strategies if necessary. This aligns with the principle of developing test instruments that are continuously evaluated and adjusted to student needs and learning objectives to be achieved.

The Model Fit Test conducted in this study provides a comprehensive evaluation of the proposed model's alignment with the observed data. Although the Chi-square test reveals a low

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

p-value (0), suggesting a potential lack of perfect fit, this result must be interpreted carefully, especially given the large sample size, which can often influence the outcome of this test (Hair et al., 2019). Despite this, other fit indices strongly support the model's adequacy. Specifically, the Comparative Fit Index (CFI) and the Tucker-Lewis Index (TLI) exceed the recommended threshold of 0.90, indicating a strong fit, while the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR) fall within acceptable limits (≤ 0.08). These indices suggest that the model adequately represents the observed phenomenon.

The model also demonstrates significant statistical findings, particularly with the latent variable "Wave," which shows a positive and statistically significant relationship with the observed latent variable, supported by a high z-value of 4.341 and a p-value close to zero. This result confirms the reliability of the wave variable's impact and positions this study as an important contribution to the field. Compared to similar studies, such as those by (Kastriti et al., 2022) and (Kaniawati et al., 2019), which focused on other physics concepts, this research's specific focus on wave phenomena provides a novel contribution to the diagnostic tool development in science education. The study's rigorous statistical validation adds to the growing body of research that aims to create reliable models for diagnosing student misconceptions in complex scientific topics, thus setting a strong foundation for future studies in other areas of science education. In sum, while the Chi-square test presents some concerns, the overall model fit and the significant findings from the parameter estimations and effect sizes affirm the robustness and relevance of the model in explaining the relationship between the latent variables and the observed data.

Using diagnostic test instruments designed with strong content validity and grounded in educational theory and prior research enables more accurate identification of students' conceptual errors. This contributes significantly to improving learning outcomes and enhancing the overall quality of physics education. Theoretically, this instrument's development enriches the educational assessment field by offering a systematic and empirically tested model for measuring student competence in physics subjects often perceived as abstract and challenging. It also reinforces key constructs in educational measurement theory, such as reliability, validity, and measurement precision, thereby raising standards in educational testing, particularly in science subjects.

From a practical standpoint, this diagnostic instrument provides tangible benefits for classroom instruction. It equips teachers with precise diagnostic information on students' learning difficulties, enabling them to tailor their teaching methods, implement targeted interventions, and deliver more meaningful feedback to support conceptual understanding. This study contributes to future scientific developments by laying the groundwork for adaptive assessment models that can respond dynamically to students' individual learning needs. As educational technology and data analytics continue to evolve, there is significant potential to integrate diagnostic tools like this into digital platforms, allowing real-time feedback and personalized learning paths. The trend toward differentiated and formative assessment practices in STEM education underscores the importance of reliable diagnostic instruments. Future research can explore the integration of such tools into learning management systems, expand their use across diverse student populations, and refine them through machine learning to enhance predictive validity and instructional relevance.

CONCLUSION

Based on the results of the analysis, it can be concluded that the diagnostic test instrument developed on the wave material in physics learning shows strong psychometric quality in terms of both validity and reliability. Content validation involving seven expert validators produced

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

an average Aiken's V score of 0.849, which is included in the "very valid" category, indicating that the items are aligned with the learning indicators and are formulated. Construct validity testing through KMO (0.946) and Bartlett's Test of Sphericity confirmed the suitability of the data for factor analysis, supported by a high average MSA value of 0.946. Furthermore, confirmatory factor analysis showed that the model adequately fits the data. Fit indices such as CFI (0.948) and TLI (0.935) exceeded the recommended threshold, while the RMSEA (0.072) and SRMR (0.043) values also indicated acceptable model fit. Although the Chi-square test yielded a low p-value (common in large samples), the model parameters indicated significant relationships between the latent constructs and the observed variables. For example, the latent variable "Wave" showed a significant effect with an estimated 0.232 and a very small p-value.

Despite these positive findings, several limitations need to be noted. The validation process involved a small number of experts, and the sample size for construct validation may affect the generalizability of the results. Future research should expand the number and diversity of respondents and apply more complex modelling techniques to validate further and refine the instrument. Including additional external variables may also increase the comprehensiveness of the evaluation and the usefulness of the diagnostic test in identifying misconceptions and guiding instruction.

Conflict of interests

There are no known conflicts of interest associated with this publication.

REFERENCES

Abdillah, M. B., Rahardjo, T. J., Martono, & Prihatin, T. (2021). The influence of transformational leadership, organizational culture, and rewards through an commitment to integrated goal to the performance of private higher education lecturers in the central java region. *Proceeding ISET (2021) Universitas Negeri Semarang International Conference on Science, Education and Technology*, *7*(1), 812–816. http://ezproxy.umgc.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=160601116&site=eds-live&scope=site

AM, M. A., & Istiyono, E. (2022). Multiple representation ability of high school students in physics: A study of modern response theory. *Thabiea: Journal of Natural Science Teaching*, *5*(1), 85–97. https://doi.org/10.21043/thabiea.v5i1.12550

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

Blessing, A., Emmanuel, N., & Chinyere, E. (2021). Effect of differentiated instruction on students' achievement in geometry. *International Journal for Research in Applied Sciences and Biotechnology*, *8*(3), 6–12. https://doi.org/https://doi.org/10.31033/ijrasb.8.3.2

Burak, D., & Gültekin, M. (2021). Verbal-visual learning styles scale: Developing a scale for primary school students. *International Journal on Social and Education Sciences*, *3*(2), 287–303. https://doi.org/10.46328/ijonses.171

Burkholder, E., Wang, K., & Wieman, C. (2021). Validated diagnostic test for introductory physics course placement. *Physical Review Physics Education Research*, *17*(1), 10127. https://doi.org/10.1103/PhysRevPhysEducRes.17.010127

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

Creswell, J. W. (2022). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

Hadi, S., Haryanto, H., AM, M. A., Marlina, M., & Rahim, A. (2022). Developing classroom assessment tool using learning management system-based computerized adaptive test in vocational high schools. *Journal of Education Research and Evaluation*, *6*(1), 143–155. https://doi.org/10.23887/jere.v6i1.35630

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2017). Multivariate data analysis. In *Pearson* (Seventh). https://doi.org/10.1002/9781118895238.ch8

Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (Eighth). Annabel Ainscow.

Haryanto, H., Kholis, N., Hadi, S., Apino, E., Asriadi AM, M., & Trizkyana, C. K. (2023). Determinant factors affecting the research performance of lecturers receiving external funds. *Research and Evaluation in Education*, *9*(2), 198–209. https://doi.org/10.21831/reid.v9i2.68457

Helmi, S., Sofyan, Y., Chantika, T., & Muh Asriadi, A. M. (2025). Linking Culture to Performance: How Organizational Culture Drives Employee Success at PT Cyberindo Aditama (CBN). *International Research Journal of Multidisciplinary Scope*, *6*(1), 441–450. https://doi.org/10.47857/irjms.2025.v06i01.02304

Homjan, S., Sri-ngan, K., & Homjan, W. (2022). Construction of diagnostic test in mathematics on addition and subtraction basic for primary students. *Journal of Education and Learning*, *11*(3), 88–94. https://doi.org/10.5539/jel.v11n3p88

Hyland, D., & O'Shea, A. (2022). The nature and prevalence of diagnostic testing in mathematics at tertiary-level in Ireland. *Teaching Mathematics and Its Applications*, *41*(1), 32–50. https://doi.org/10.1093/teamat/hrab006

Istiyono, E. (2022). Diagnostic tests as an important pillar in today's physics learning: Four-tier diagnostic test a comprehensive diagnostic test solution. *Journal of Physics: Conference Series*, *2392*(1), 0–8. https://doi.org/10.1088/1742-6596/2392/1/012001

Istiyono, E., Sunu, W., Fenditasari, K., Rai, M., & Saepuzaman, D. (2023). The development of a four-tier diagnostic test based on modern test theory in physics education. *European Journal of Educational Research*, *12*(1), 371–385. https://doi.org/10.12973/eu-jer.12.1.371

Johnson, R. B., & Christensen, L. (2017). Educational research: Quantitative, qualitative, and mixed approaches. In *SAGE Publications, Inc.*

Kaniawati, I., Fratiwi, N. J., Danawan, A., Suyana, I., Samsudin, A., & Suhendi, E. (2019). Analyzing students' misconceptions about newton's laws through four-tier newtonian test (FTNT). *Journal of Turkish Science Education*, *16*(1), 110–122. https://doi.org/10.12973/tused.10269a

Kastriti, E., Kalogiannakis, M., Psycharis, S., & Vavougios, D. (2022). The teaching of Natural Sciences in kindergarten based on the principles of STEM and STEAM approach. *Advances in Mobile Learning Educational Research*, *2*(1), 268–277. https://doi.org/10.25082/amler.2022.01.011

Martiana, A., Istiyono, E., & Widihastuti, W. (2022). Critical sociology in the development of HOTS-oriented cognitive assessment instruments. *Journal of Social Studies (JSS)*, *18*(2), 197–206. https://doi.org/10.21831/jss.v18i2.51430

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

10.21831/pep.v29i1.84655

Navarro, D., Foxcroft, D., & Faulkenberry, T. J. (2019). *Learning stafisticks with JASP: A tutorial for psychology students and other beginners*. (Version 1/2).

Rahman, K. A., Hasan, M. K., Namaziandost, E., & Ibna Seraj, P. M. (2021). Implementing a formative assessment model at the secondary schools: attitudes and challenges. *Language Testing in Asia*, *11*(1), 1–18. https://doi.org/10.1186/s40468-021-00136-3

Scoulas, J. M., Aksu Dunya, B., & De Groote, S. L. (2021). Validating students' library experience survey using rasch model. *Library and Information Science Research*, *43*(1), 101071. https://doi.org/10.1016/j.lisr.2021.101071

Shanmugam, S. K. S. (2018). Determining gender differential item functioning for mathematics in coeducational school culture. *Malaysian Journal of Learning and Instruction*, *15*(2), 83–109. https://doi.org/10.32890/mjli2018.15.2.4

Spiller, L., & Tuten, T. (2019). Assessing the pedagogical value of branded digital marketing certification programs. *Journal of Marketing Education*, *41*(2), 77–90. https://doi.org/10.1177/0273475318822686

Tabatabaee-Yazdi, M., Motallebzadeh, K., Ashraf, H., & Baghaei, P. (2018). Development and validation of a teacher success questionnaire using the rasch model. *International Journal of Instruction*, *11*(2), 129–144. https://doi.org/10.12973/iji.2018.11210a

Tomczyk, Ł., & Eger, L. (2020). Online safety as a new component of digital literacy for young people. *Integration of Education*, *24*(2), 172–184. https://doi.org/10.15507/1991-9468.099.024.202002.172-184

Watkins, M. W. (2018). Exploratory factor analysis: A guide to best practice. *Journal of Black Psychology*, *44*(3), 219–246. https://doi.org/10.1177/0095798418771807

ISSN (print) 2685-7111 | ISSN (online) 2338-6061