

Research Article

Cite this article: Lamanepa, G. H., Istiyono, E., Rosnawati, R., Arsisari, A., & Hali, F. (2026). Construct Validity of Computer-Based Physics Test (CBPsyT) Through Confirmatory and Item Response Theory: A Complementary Approach to Test Development. *Educational Process: International Journal*, 22, e2026054. <https://doi.org/10.22521/edupij.2026.22.54>

Received October 10, 2025

Accepted December 3, 2025

Keywords: Construct validity, CFA, IRT, CBPsyT, multiple representation test, physics construct

Author for correspondence:

Godelfridus Hadung Lamanepa



godelfridushadung.2022@student.uny.ac.id

Universitas Negeri Yogyakarta, Indonesia



OPEN ACCESS

© The Author(s), 2026. This is an Open Access article, distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction, provided the original article is properly cited.

Construct Validity of Computer-Based Physics Test (CBPsyT) Through Confirmatory and Item Response Theory: A Complementary Approach to Test Development

Godelfridus Hadung Lamanepa^{ID}, Edi Istiyono^{ID}, Raden Rosnawati^{ID}, Ayen Arsisari^{ID}, Fitriyani Hali^{ID}

Abstract

Background/purpose. This study aims to validate the computer-based physics test (CBPsyT) construct through confirmatory factor analysis (CFA) to ensure that CBPsyT is suitable for its intended use. However, CFA studies are based on weak theoretical perspectives, a lack of testing of alternative theoretical views, or insufficient evidence for construct validity, so item response theory (IRT) analysis needs to be added to this study.

Materials/methods. The research is exploratory and includes a quantitative approach. We developed CBPsyT according to test development procedures consisting of twelve steps and administered it to 516 students. Construct validity analysis based on the CFA and IRT perspectives using the R program.

Results. The relationship between indicators and latent constructs, with multiple representations in CBPsyT, is indicated by indicator loadings > 0.70, which denote a strong relationship between the two. Internal consistency is demonstrated by a composite reliability > 0.80, and CFA model fit values indicate that the hypothesized model is consistent with the empirical data. The IRT results add that discrimination and difficulty scores range from -2 to +2. The test information function indicated that the instrument was reliable for ability (Theta) scores ranging from -3 to +3.

Conclusion. The test quality profile was complete when CFA and IRT were combined. To the best of our knowledge, this article provides practical information on the psychometric characteristics of CBPsyT and guides new researchers in demonstrating construct validity.

1. Introduction

Estimating the construct of a test instrument in a measurement is important because the results reveal the true nature of the instrument in question (Weyers et al., 2023). Construct validity is one type of validity, similar to content validity and predictive validity, and is used to assess the extent to which a research instrument or test actually measures the abstract concept or construct to be calculated in accordance with the research objectives (Ohiri et al., 2024). Without construct validity, research risks mismeasuring other irrelevant concepts, resulting in erroneous conclusions and lowering the quality of the research (Malone et al., 2021). Several previous researchers (Jiang et al., 2023; Phanniphong & Na-Nan, 2025) confirmed the importance of construct validity through factor analysis, but other researchers (Bean & Bowen, 2021; Thissen, 2022) argue that factor analysis needs to be supported by IRT to provide more in-depth information about the quality and difficulty level of items and the ability of respondents, both of which complement each other for a comprehensive evaluation of the instrument.

The CBPsyT is a computer-based measurement tool used to assess the multiple-representation abilities of high school physics students. This instrument was developed using the test development steps outlined by Kyriazos and Stalikas (2018). In addition to evidence of content validity from experts, the CBPsyT requires construct validity to provide more meaningful information than content validity alone. Construct validity via CFA is intended to test whether the developed measurement tool actually measures the desired construct, based on existing theories or hypotheses (Kang et al., 2025). Test outcomes can be shown through regression weights, standardized regression weights, convergent validity, extracted variance, construct reliability, and discriminant validity (Kang et al., 2025; Said et al., 2011). However, CFA studies based on weak theoretical foundations, a lack of testing alternative theoretical perspectives, or limited evidence may not adequately support construct validity (DiStefano & Hess, 2005). CFA ensures that test data conform to predetermined theoretical constructs, thus improving data validity, while IRT focuses on the relationship between item performance and respondent ability (Kang et al., 2025). IRT offers detailed insights into item discrimination and difficulty when measuring constructs, providing comprehensive information about items and test performance. The CFA and IRT analyses in this study complement each other by providing unique and supporting information about the quality of the CBPsyT we developed. Combining these methods offers strong evidence of validity, which is crucial for effective research and practice.

The function of IRT in developing measurement tools is to analyze item characteristics and respondents' latent traits to create more valid and reliable assessments (Yang & Kao, 2014). IRT helps select high-quality items, design adaptive tests, and ensure that measurement tools perform well across different levels of respondent ability, leading to more accurate and fair results. The key parameters in IRT are item discrimination, item difficulty, and guessing factors (Min & Aryadoust, 2021). The discrimination parameter (a) shows an item's ability to distinguish individuals at different trait levels. The difficulty parameter (b) measures the challenge of an item, with higher values indicating greater difficulty. The guessing parameter reflects the probability of guessing the correct answer, which is especially relevant for multiple-choice questions. IRT offers valuable insights by assessing a person's ability more precisely through analysis of item features such as difficulty, rather than relying solely on the number of correct answers. Additionally, IRT can automatically adapt questions to match participants' abilities, helping test developers understand and improve question quality.

The Item Response Theory model has been widely applied in education to understand the relationship between responses to test items and the underlying latent ability of test takers (Garcia et al., 2018). Item Response Theory analysis enhances construct analysis in CFA by providing more detailed information about individual item performance and validating construct validity through

reliability and measurement invariance checks across different groups, which CFA alone cannot measure (Kamata & Bauer, 2008). IRT also helps test the assumptions underlying CFA and supplements its analysis with a method that is more sensitive to scale quality, especially in short scales. IRT and CFA work together because CFA is used to confirm the structure of psychometric models. At the same time, IRT offers detailed insights into the performance of each test item, resulting in a comprehensive view of assessment quality and the constructs being measured (Kim & Yoon, 2011). CFA confirms the presence of latent constructs, and IRT explains how specific items within these constructs function, forming the basis for test development and refinement. This study assesses the construct validity of computer-based physics tests (CBPsyT) while analyzing them from an IRT perspective to gain a more complete understanding of the constructs and gather information for test development.

Multiple representation constructs in physics are a learning approach that utilizes various formats to present the same physics concept, such as verbal explanations, diagrams, graphs, and mathematical equations. This approach is crucial because it enables students to develop a deep understanding and overcome difficulties in solving physics problems by connecting different perspectives on a concept. Multi-representation is important for teachers to assess because it helps students build a deep understanding of concepts, improves problem-solving skills, and complements the experience gained from a single representation. By assessing students' multi-representation skills, teachers can determine how well students integrate various formats, such as diagrams, graphs, equations, and words, to gain a more comprehensive understanding of science and mathematics concepts.

2. Literature Review

Multiple representation skills are necessary in physics, given the structure of physics, whose concepts are abstract and often related to complex natural phenomena. One suitable learning strategy in physics is multiple representation-based learning. Almost all subjects offer various representations, such as photographic images used as text illustrations, and explanatory diagrams that provide meaning to the text (de Jong & van der Meij, 2012). Multiple representation skills refer to the ability to present or reapply the same concept in different formats, including images, diagrams, graphs, and verbal and mathematical representations (Treagust & Duit, 2017). Multiple representations are used to complement each other, as a single representation is often insufficient to convey all the necessary information for learning physics. The implication is that multiple representations are used in teaching to promote learning.

Multiple representations are categorized into two main types: quantitative representations, which include mathematical representations, and qualitative representations, which include images, graphs, and diagrams (Treagust & Duit, 2017). Experts often use qualitative representations, such as images, graphs, and diagrams, to help them understand problems before using mathematical equations and solving them quantitatively. Diagrammatic representations are more concrete and often used to facilitate understanding of abstract concepts, whereas mathematical representations are more necessary when solving quantitative problems (Ofosu et al., 2020). The ability to use multiple representations, including mathematical, verbal, diagrammatic, and graphical representations, is essential when studying physics. The goal is not only to complement each other but also to encourage students to build a deeper understanding of a situation (Carpenter et al., 2020; Yoon et al., 2021). The primary reason for using multiple representations is to leverage the diverse learning processes that different representations facilitate.

The objectives and functions of assessment are fulfilled when the measurement tool possesses strong psychometric properties. Psychometric properties refer to characteristics that evaluate a measurement tool's reliability and validity. Reliability indicates the consistency and stability of the

tool, while validity measures its accuracy and precision in assessing the intended construct (Allen & Yen, 1979; Taherdoost, 2018). Validity is classified into types: (1) content validity, (2) criterion validity (both concurrent and predictive), and (3) construct validity (including discriminant and convergent validity) (Cohen & Swerdlik, 2018). Validity is a unified concept encompassing content, criterion, and construct validity. Evidence supporting validity is not presented separately but integrated comprehensively and aligned with the test's purpose.

Factor analysis, a series of psychometric tests used to examine the dimensions (internal structure) of a measurement construct, is perhaps the most commonly used method for testing construct validity in social science research (Faller et al., 2006; Husain & Aziz, 2022; Orcan, 2018). Factor analysis offers insights into reliability, item quality, and construct validity. Its main goal is to determine whether and how well the items on a scale can represent the underlying hypothesis of the construct or constructs, known as factors. It is also a highly sensitive analytical method for identifying problematic items and deciding the number of factors (Beavers et al., 2013). Confirmatory Factor Analysis (CFA) is a "theory testing strategy" that relies on measurement modeling to assess whether the factor structure of an existing measure remains consistent when applied to a new sample of participants (Plucker, 2003; Walton et al., 2023).

In confirmatory factor analysis, researchers first formulate hypotheses about the underlying factors of the measurements they use, and they can impose constraints on the model based on these prior hypotheses (Arifin & Yusoff, 2016). Confirmatory Factor Analysis (CFA) provides information on how well a theoretical model describing the relationships between latent constructs and their indicators fits the actual data (Ghazali & Nordin, 2019). This information includes an overview of the model, such as which indicators relate to each latent factor and the correlations among factors, as well as the theoretical and empirical support for the model. The results also include the instrument's validity and reliability, such as factor loading values, correlation coefficients, Average Variance Extracted (AVE), and model fit indices. Through CFA, we can measure whether the items consistently represent the construct, ensuring the construct validity and reliability of the measurement instrument (Herwin & Nurhayati, 2021).

Nowadays, experts and researchers refer to item response theory as the criterion for the validity and reliability of a test or other measurement instruments. Several psychometric experts, such as DeMars (2008), Hambleton & Jones (1993b), and Lord (1953), are interested in the idea that measurement practices would be improved if test items and statistics could be treated as independent samples, which would lead to more accurate measurements. The meaning of validity in IRT indicates that the characteristics of test items are not influenced by the sample from which they are estimated. Even if the same items are given to different groups, they will produce the same scores and rankings. The Item information function estimates reliability coefficients in IRT for dichotomous and polytomous data. Item response models are also referred to as robust models because their underlying assumptions are highly rigorous and therefore unlikely to be satisfied by test data.

Item response theory uses a fundamentally different approach from classical test theory. Item response theory is a non-linear model that provides the probability of a correct response to an item as a function of item characteristics and test taker ability (Embretson & Reise, 2000: 20). Item response theory is based on two main postulates, namely: (1) test takers' performance on a test item can be predicted or explained by a set of hidden factors, or abilities, and (2) the relationship between test takers' performance on an item and the characteristics underlying item performance can be described by an item characteristic curve (Hambleton et al., 1992: 7). This distinguishes item response theory from classical test theory.

The information function of items and tests in IRT is useful for determining the accuracy of measurement at a certain ability level and expressing the strength or contribution of test items in

revealing the latent trait measured by the test (Hambleton & Jones, 1993). The Item information function assists in test development by identifying which items are suitable for the model, thereby aiding in test item selection. This information function is important to consider in test development because the greater the information function, the better the test item. The information function has an inverse relationship with measurement error. The greater the item information for a given ability level, the less error is involved in estimating the ability of test takers (Hambleton & Jones, 1993: 42). This means that the more tests are given at a certain ability level, the smaller the error associated with the estimated ability, and vice versa.

3. Methodology

3.1. Procedures

The research is exploratory and includes a quantitative approach. This study uses adaptation instruments in the context of development research. The procedure for developing the adaptation instrument is based on the test instrument development model. This process consists of eleven important stages, including determining test objectives, descriptions, and specifications; operational definitions; scoring; expert item reviews; test prototypes; determining test subjects; readability tests; experts' revisions; empirical trials; and analyzing test characteristics. These stages have been undertaken to develop a valid and reliable instrument to assess high school students' physics skills.

3.2. Participants

This quantitative study involved 516 high school students in Kupang, East Nusa Tenggara Province, Indonesia. All students who took the test were grade XI students from six different schools. The students were selected considering comparable background materials, test construction, and topics covered. The convenience sampling method was used in this study due to the large population size and the researcher's inability to obtain a representative random sample from the accessible population. This study utilized a comprehensive assessment to evaluate the validity and reliability of the developed instrument. Data were collected through a computer-based test that generated answers from 516 participants from six different high schools. All participants in this study had provided informed consent through a written permission form included at the beginning of the study.

3.3. Instrument

The test instrument used in this study is a computer-based physics test (CBPsyT) consisting of 20 multiple-choice questions. The physics material covered in this test is mechanics.

3.4. Data Collection and Model

To achieve the objectives of this study, we conducted computer-based physics tests on 516 high school students. The following section covers these methods and their respective results. Hypothetically, the CBPsyT model comprises four aspects: verbal representation (RV), diagram representation (RD), graphic representation (RG), and mathematical representation (RM).

Table 1. Aspects and indicators of the CBPsyT instrument

Aspect	Indicator	Item Code
Verbal representation (RV)	Constructing interpretations of physical principles or laws verbally or in writing from mathematical forms, images, or symbols	RV1, RV2
	Connecting physics concepts to everyday life verbally or in writing about physics issues	R3, RV4

	Constructing metaphors or analogies by connecting similar concepts to things that are often encountered or experienced in everyday life	RV5, R6
Diagram representation (RD)	Creating a free-body diagram or other types of diagrams	RD1, RD2, RD3
	Grouping the various forces acting within a system	RD4, RD5, RD6
Graphic representation (RG)	Identifying the data presented on the graph	RG1, RG2
	Review the results based on the conclusions obtained through the graph.	RG3, RG4
Mathematical representation (RM)	Create relevant equations based on physical system considerations	RM1, RM 2
	Predicting the implications of mathematical formulas on changes in an object or physical principle	RM3, RM4

3.5. Data Analysis Techniques

The analysis was conducted in two stages. First, CFA was used to confirm the construct validity of the identified factor structure. The following are the steps required to demonstrate CFA: (1) checking the adequacy of the KMO sample, (2) defining the construct and measurement model. (3) specifying the number of factors and loading patterns based on theory, ensuring that the model can be identified, which means that all parameters can be estimated with certainty. (4) estimate the model parameters: use estimation techniques such as Maximum Likelihood Estimation (MLE) or the least squares method to estimate the relationship between latent and manifest variables. CFA construct validity analysis includes convergent validity tests to ensure that the items measure the same construct and discriminant validity tests to ensure that different constructs are indeed distinct. Both tests involve evaluating factor loadings and Average Variance Extracted (AVE) values, as well as comparing AVE with squared correlations between constructs. (5) assessing model fit: examining the fit of the model to the observed data using various goodness-of-fit indices.

The second step is to use IRT to analyze item characteristics, such as difficulty and discrimination levels, and to test information functions to measure the extent to which the test can distinguish among participants' abilities within a given ability range. The following are the steps in IRT analysis: (1) Checking the suitability of the IRT model used: ANOVA is used to select the IRT model that best fits the available data, for example, choosing between one, two, or three parameter models by looking at certain statistical values such as the Akaike information criterion (AIC). (2) IRT analysis according to the selected model: IRT analysis in this study uses IRT 2PL. This model predicts the probability of a person answering an item correctly based on two parameters: the item discrimination parameter and the difficulty parameter. (3) Item information function (IIF) and test information function (TIF) analysis: IIF is used to analyze the characteristics of each item individually (difficulty and discrimination power), while TIF is used to evaluate the overall measurement information of the test within a certain range of participant abilities, as well as the overall reliability and validity of the test.

4. Results

4.1. CFA Results

Before conducting further analysis of the test items and instruments, a feasibility test was first carried out, using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) and Bartlett's

test. KMO is a test used to examine the strength of the partial correlation (how the factors explain each other) between variables. KMO values closer to 1.0 are considered ideal, while values less than 0.5 are unacceptable. Recently, most scholars argue that a KMO of at least 0.80 is good enough for factor analysis to commence (Astuti et al., 2024). KMO and Bartlett's test assess sampling suitability, which means whether the responses obtained from the sample are adequate. Bartlett's test of sphericity is a statistical test used to assess the hypothesis that the variables are not correlated in the population, or to assess the homogeneity of the data. Bartlett's statistic is designed to test for equality of variances across groups against the alternative that variances are unequal for at least two groups (Vorapongsathorn et al., 2004). A significant Bartlett's test is indicated by $p < 0.05$ (Lovric, 2011). MSA is an index comparing partial correlation coefficients for each variable. The MSA test is used to measure sample adequacy.

Table 2. Bartlett's & KMO Test

Kaiser Meyer Olkin Measure of Sampling Adequacy		0.85
Bartlett's test of Sphericity	approx. chi square	199.59
	df	19
	p-value	< 2.2e-16

The KMO value in Table 2 is 0.85, indicating substantial information overlap among the variables, suggesting a strong partial correlation. Hence, it is plausible to conduct factor analysis. Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is the identity matrix. An identity correlation matrix indicated that your variables are unrelated and are not ideal for factor analysis (Odoi et al., 2022). A significant statistical test, typically with a p-value less than 0.05, indicates that the correlation matrix is not an identity matrix (rejection of the null hypothesis), as shown in Table 2.

In this study, CFA was employed to evaluate the extent to which the observed variables accurately reflected the latent variables, specifically multiple representation abilities. CFA was also used in this study to test whether the measures of the construct were consistent with the researchers' understanding of its nature. The validity of the test instrument construct comprises four latent variables, namely RV, RD, RG, and RM, along with their respective indicators. The results of the CFA construct test analysis (Figure 1) show factor loading values, indicating the strength and direction of the relationships between the indicators and the latent variables. Referring to the standardized factor loading values (normalized scale) of each indicator, RV1, RV2, RV3, RV4, RV5, and RV6, respectively, provide an overview of the overall strength of the relationship with the RV variable, which is 0.85, 0.85, 0.87, 0.92, 0.98, and 0.98. Overall results of factor loading contributions for latent variables.

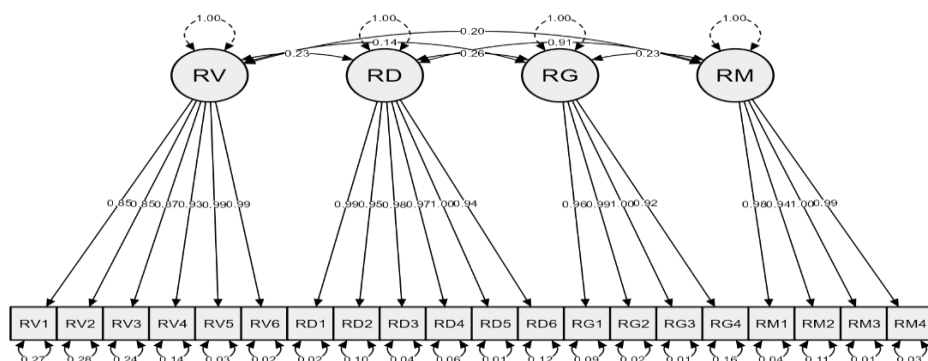


Figure 1. The CBPsyT construct model in CFA

The results of the CFA construct test analysis (Figure 1) require interpretation of several aspects, including: first, the factor loading in the CFA construct test measures the strength and direction of the relationship between each indicator and the latent variable construct being measured. The factor loading values for each indicator of each latent variable meet the criteria as indicated by Hair et al. (2013). According to their guidelines, a factor loading greater than 0.70 indicates that the indicator reflects or represents the intended construct. Second, in CFA testing, factor loadings and AVE values are two distinct yet interrelated metrics used to evaluate the convergent validity of the measurement model. Factor loadings indicate the strength of the relationship between the indicator and the latent factor being measured. In contrast, AVE values measure the proportion of variance that is extracted by the latent factor from the indicator, relative to the variance due to error. Table 3 below shows the AVE values from the CFA construct test results.

Table 3. Average Variance Extracted

Factor Average variance extracted (AVE)	
RV	0,831
RD	0,942
RG	0,933
RM	0,955

The AVE values (Table 3) represent the average variance explained by each latent factor listed above. Each value indicates the measure of convergent validity of the construct formed from this analysis. An AVE value greater than 0.50 (Hair et al., 2013) suggests that the construct explains more variance than its indicators, indicating good convergent validity.

This study also considers discriminant validity, which refers to the extent to which different constructs in the research model are truly distinct from one another. This ensures that the indicators used to measure a construct correlate more strongly with that construct than with other unrelated constructs. The HTMT value (Table 4) was analyzed as a measure of discriminant validity to determine the average correlation between different indicators of two constructs (heterotrait-heteromethod correlation), compared to the average correlation between indicators measuring the same construct (monotrait-heteromethod correlation).

Table 4. Heterotrait-Monotrait Ratio

	Heterotrait-monotrait Ratio			
	RV	RD	RG	RM
RV	1.000			
RD	0.208	1.000		
RG	0.124	0.268	1.000	
RM	0.190	0.922	0.244	1.000

In this analysis, discriminant validity of HTMT is used to distinguish among different constructs. The results (Table 4) show that the discriminant validity between the RD variable and the RV variable is 0.208, while that between the RG variable and the RV variable is 0.124. An HTMT value greater than 0.85 (Hair et al., 2013) indicates good discriminant validity, meaning that the measured constructs are distinct from one another.

Along with factor loading and AVE analysis, CFA analysis also requires fit index values (Table 5) to evaluate how well the proposed measurement model fits the empirical data.

Table 5. Fit Index Measurement Model

Fit Indicates				
	Index	Results	Cut off index	Description
	Comparative Fit Index (CFI)	0.92	CFI > 0,87	Fit
	Tucker-Lewis Index (TLI)	0.94	TLI > 0,90	Fit
	Bentler-Bonett Non-normed Fit Index (NNFI)	0.92	NNFI > 0,90	Fit
	Bentler-Bonett Normed Fit Index (NFI)	0.92	NFI > 0,90	Fit
	Bollen's Relative Fit Index (RFI)	0.91	RFI > 0,90	Fit
	Bollen's Incremental Fit Index (IFI)	0.91	IFI > 0,90	Fit
	Root mean square error of approximation (RMSEA)	0.071	RMSEA < 0,08	Fit
	Standardized root mean square residual (SRMR)	0.010	SRMR < 0,10	Fit

The values of each model fit index reported in the model fit test meet the minimum requirements for an appropriate measurement model. The model fit measures obtained assess how well the proposed model captures the covariance between all items or measures in the model. Thus, the purpose of confirmatory factor analysis is to test whether the data fit the hypothesized measurement model. This hypothesized model is based on theory or previous analytical research.

The reliability of the composite omega (Table 5) indicates how well the variables underlying a construct are represented by the test items in the research instrument, with a value of $\omega > 0.70$ indicating adequate representation.

Table 5. Alpha & Omega Reliability Coefficient

Reliability		
	Coefficient ω	Coefficient α
RV	0,956	0,970
RD	0,989	0,990
RG	0,979	0,983
RM	0,989	0,988
Total	0,989	0,939

The results of the omega reliability analysis, with values greater than 0.70 on the RV, RD, RG, and RM aspects, indicate that the research instrument developed has good internal consistency and is reliable for consistently measuring the same construct. A coefficient value greater than 0.70 suggests that most of the observed score variation is due to the true score, and only a small portion is caused by measurement error. The omega and alpha reliability values differ because omega is based on a factor analysis model that estimates the variance of test scores explained by latent factors. In contrast, alpha is based on inter-item correlations, which assume that all items measure the same construct equally. Because of these differences in underlying assumptions, omega can be a more accurate estimate of reliability, especially when there are violations of the item-equivalence assumption in alpha.

4.2. Characteristics of CBPsyT according to IRT

IRT analysis begins with item fit analysis, a type of data-model fit evaluation specifically designed for test-item performance. This analysis is crucial for interpreting and understanding test results and evaluating item performance. In IRT, item fit analysis is related to the use of logistic functions. Model selection criteria can also be determined using Akaike's information criterion (AIC) (Akaike, 1974), a technique based on the sample fit to estimate the likelihood of a model in predicting future values. A good model has the minimum AIC among other models. AIC can be used to choose between additive and multiplicative Holt-Winters models.

Table 6. Anova Fit Model

	AIC	BIC	Log. lik	df	p. value
Anova					
1PL	14044.43	14155.33	-6996.22		
2PL	13754.43	13967.69	-6827.21	24	<0.001
Anova					
2PL	13754.43	13967.69	-6827.21		
3PL	13802.82	13978.72	-6986.41	25	<0.001

The AIC value of the 2PL model (Table 6) is lower than those of the 1PL and 3PL models. This indicates that the 2PL parameter model provides a better fit to the data than the 1PL and 3PL models, suggesting that the 2PL model is a more suitable fit than other logistic models.

Based on model fit testing, the 2PL model was the most appropriate for data analysis, as all items were well-fitted to it. The next analysis estimated the item parameter values based on the 2PL model (see Table 7), specifically the difficulty and discrimination parameters. An item is considered good if its difficulty is in the range of -2 to +2 and its discrimination index is in the range of 0 to 2 (Hambleton & Slater, 1997).

Table 7. Difficulty & Discrimination Index of the CBPsyT Instrument

	Discrimination	Difficulty	Description
Item 1	0.826	0.442	Good
Item 2	1.033	1.621	Good
Item 3	1.470	0.013	Good
Item 4	0.243	1.926	Good
Item 5	1.677	1.015	Good
Item 6	1.486	0.600	Good
Item 7	1.482	-2.261	Good
Item 8	1.276	-1.049	Good
Item 9	1.601	1.366	Good
Item 10	1.180	-0.536	Good
Item 11	2.529	0.235	Good
Item 12	2.479	-0.996	Good

Item 13	0.599	1.683	Good
Item 14	1.976	-0.557	Good
Item 15	2.459	0.291	Good
Item 16	0.756	0.307	Good
Item 17	1.343	-1.851	Good
Item 18	1.146	0.988	Good
Item 19	0.905	0.293	Good
Item 20	1.019	-0.397	Good

The discrimination and difficulty indices of all items are acceptable because they are categorized as good. This shows that the items developed are in accordance with the rules of item response theory. Information about test items in IRT can also be obtained from the item characteristic curve (ICC), a graphical representation that illustrates the relationship between the probability of a student answering a question correctly and their ability level. The ICC graph (Figure 2) shows how an item behaves and how effectively it differentiates students by ability.

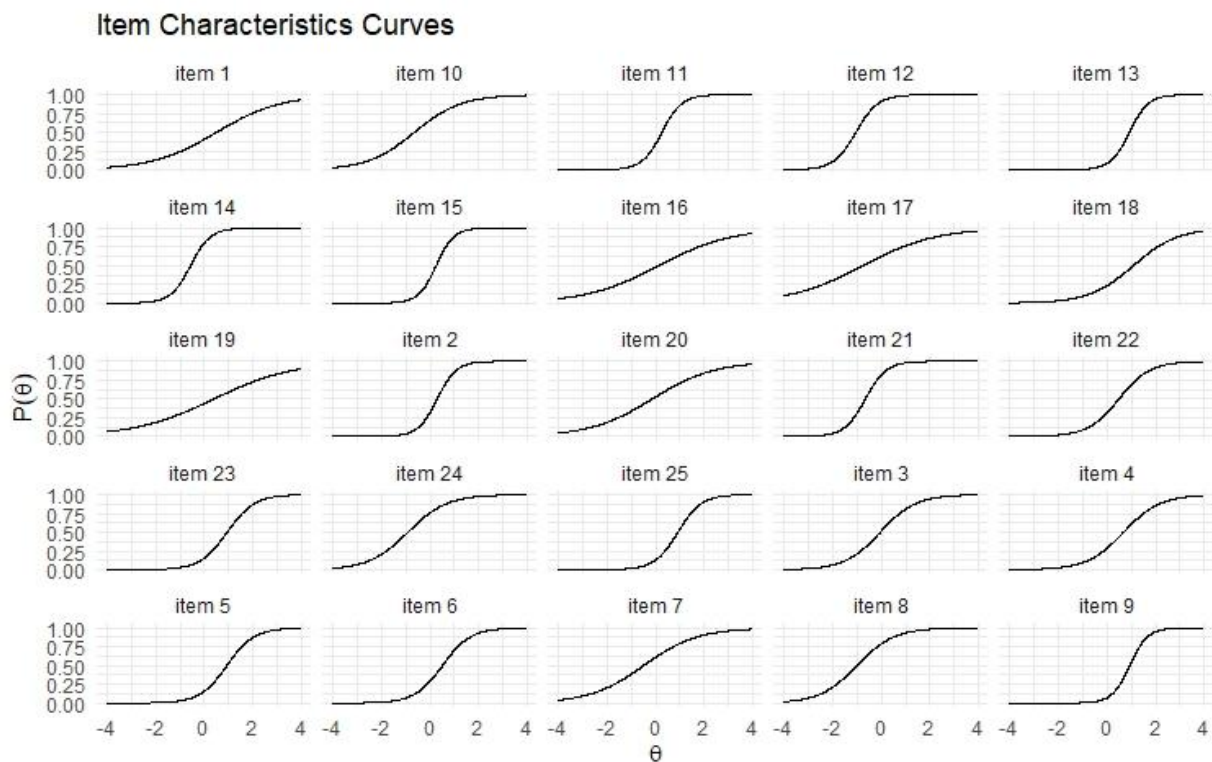


Figure 2. Item Characteristics Curves

The x-axis (see Figure 2) shows student ability. The farther to the right, the higher the individual's ability, while the y-axis shows the probability that students answered the items above correctly.

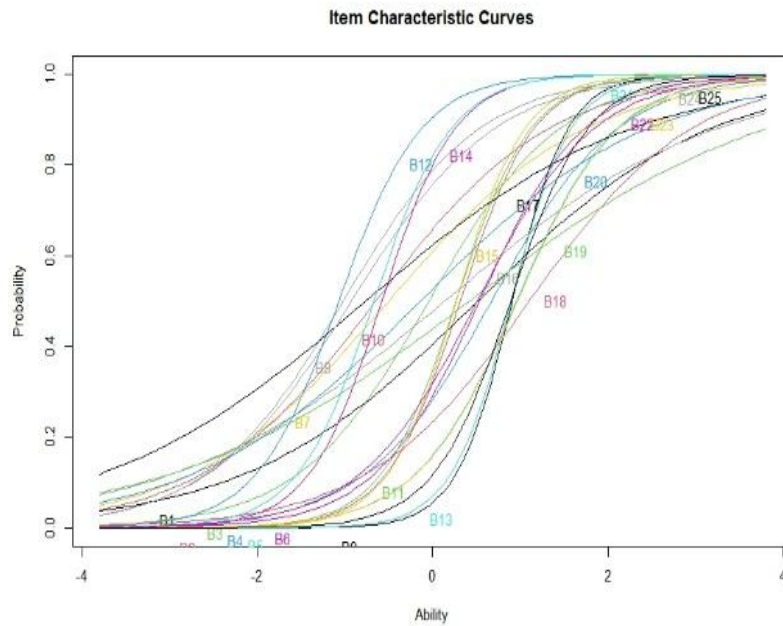


Figure 3. Item Characteristic Curves

On the item characteristic curve (see Figure 3), difficult items will shift to the right of the scale, indicating a higher ability of respondents who answer correctly. In contrast, easier items will change to the left of the ability scale. The two-parameter logistic model predicts the probability of a correct answer using two parameters. The discrimination parameter may vary between items. Furthermore, the ICCs of different items can intersect and have different slopes. The steeper the hill, the higher the item's discrimination index, as it can detect subtle differences in respondent ability.

The information function and the measurement standard deviation must be considered when selecting the model. The information function shows the extent to which each model can provide information. The higher the peak of the information function (see Figure 4), the higher the information that a model can provide.

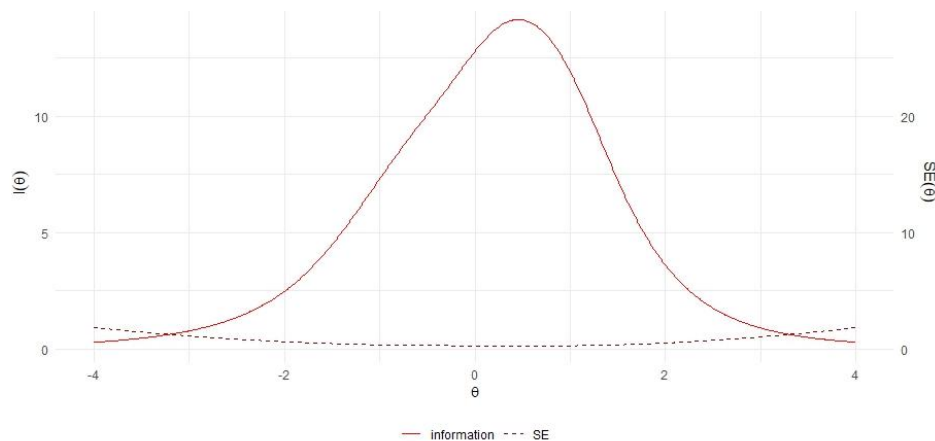


Figure 4. Test Information Function & Measurement Error

The information obtained from the measurement results is not error-free. The relationship between measurement error and information function is inversely proportional to the square, so that the greater the information function, the smaller the measurement error. The information function (see Figure 4) indicates that this developed instrument will provide reliable information when administered to students with abilities ranging from -3 to +3 logits.

5. Discussion

The CBPsyT instrument features a multi-representation construct comprising RV, RD, RG, and RM. The use of RG can summarize a large amount of information while still providing detailed information. Graphical representation is a powerful and effective tool for rendering synthetic data and representing relationships between physical quantities, as well as for data interpretation, which makes it easier (Stefanel, 2019). RD representation is significant for teaching concepts in mechanics theory, especially the concept of force, which is taught from elementary school to university. Creating diagrams is a useful step in simplifying the information presented about a problem, which will then be very helpful in solving it. RV relates to the conceptual description of a phenomenon and to a theory from a scientific perspective (Svensson & Campos, 2022). Verbal representation is intended to define, explain, analyze, and relate ideas or concepts to everyday contextual situations.

This paper explains test constructs from a confirmatory perspective and IRT, how CFA confirms the multidimensional structure of the constructs underlying test scores, and how IRT provides an in-depth understanding of the quality of individual items and scales as a whole, enabling the validity of test scores and their use in research and test development practices. CFA estimates construct validity by ensuring that indicator factor loadings are significant and by examining the goodness-of-fit indices of the CFA model. Convergent validity is evident from high factor loadings, while discriminant validity is tested by comparing AVE between constructs.

We analyzed the CBPsyT instrument using the CFA and IRT perspectives. The concept of validity in IRT relates to the extent to which participants and test items are consistent with the abilities measured by the test items. This means that any test can rank individuals according to their abilities and rank test items according to their level of difficulty (Baker & Kim, 2017; Lauwaert, 2023). In line with this objective, we present various forms of evidence to demonstrate the reasonableness of their interpretation of the arguments for validity evidence.

Factor analysis, a series of psychometric analyses used to test the internal structure and dimensions of measurement constructs, is perhaps the most widely used procedure for assessing construct validity in social science research (Faller et al., 2006; Husain & Aziz, 2022; Orcan, 2018). Factor analysis provides information about reliability, item quality, and construct validity. Its general purpose is to determine whether and to what extent items on a scale can reflect the underlying hypotheses of a construct or constructs, known as factors, and to serve as a highly sensitive analytical method for identifying problematic items and determining the number of factors.

6. Conclusion

The CBPsyT construct comprises four aspects: RV, RD, RG, and RM, consisting of a total of 20 items. The developed test construct is valid and reliable, as indicated by the indicator loading values for each aspect > 0.70 , and model fit statistics: RMSEA = 0.071, CFI = 0.92, SRMR = 0.010, NFI = 0.92, NNFI = 0.92, TLI = 0.94, CFI = 0.91, RFI = 0.91, with construct reliability > 0.80 . Confirmatory testing for the construct revealed factor loadings > 0.7 , and both convergent validity (AVE) and discriminant validity (HTMT) met the minimum criteria. Furthermore, item analysis using IRT theory with the 2PL model indicates that items with acceptable discrimination and difficulty fall within the range of -2 to +2. The test information function shows that the developed instrument is reliable when administered to students with logit abilities between -3 and +3.

This paper provides practical information on the psychometric characteristics of the developed physics test instrument. This instrument can assess students' fundamental mechanical physics skills, inform curriculum development, and enhance teaching strategies. By providing insight into these characteristics and aspects, this study also advances our understanding of this awareness as a means of estimating items in developing test instruments. In addition, this study contributes to the ongoing

discussion on the good characteristics that researchers have developed. The results of this study can inform future theoretical and practical studies and guide new researchers in developing physics test instruments.

7. Suggestion

This study has potential limitations, including the use of convenience samples in quantitative studies, which is a methodologically weak approach and only applicable to this sample; therefore, the findings cannot be generalized. Sample selection interventions are still in place due to the large area of the province of East Nusa Tenggara, Indonesia. EFA analysis may be necessary in this study, but it was not performed. The IRT model's unit of analysis is the item; items can be compared across measures, provided they measure the same latent construct. Furthermore, they can be used in differential item functioning analyses to assess why calibrated and tested items still behave differently across groups. This can lead to research to identify the causative agents behind differences in responses and to link them to group characteristics. Finally, they can be used in Computerized Adaptive Testing.

Declarations

Author Contributions. Godelfridus Hadung Lamanepa: Corresponding author, conceptualization, design, data acquisition, data analysis, drafting manuscript. Edi Istiyono: critical revision of the manuscript, reviewing. Raden Rosnawati: critical revision of the manuscript, reviewing. Ayen Arsisari: create tools for research and data analysis. Fitriyanihali: literature review, create tools for research, and data analysis.

Conflicts of Interest. The authors declare no conflict of interest.

Funding. No Funding

Ethical Approval. This research has obtained written permission from Yogyakarta State University, Number B/1086.UN34.17.LT.2025 has been approved by respondents by filling out a consent form via Google Form, as explained in the research methodology.

Data Availability Statement. The data supporting the results reported in this study are available from the corresponding author upon reasonable request.

Acknowledgments. The authors want to express their deepest gratitude to the Lembaga Pengelola Dana Pendidikan (LPDP), Pusat Pembiayaan dan Asesmen Pendidikan Tinggi (PPAPT), and Beasiswa Pendidikan Indonesia (BPI), who have sponsored our doctoral studies.

References

- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Allen, M. j., & Yen, W. M. (1979). Introduction to Measurement. In *Measurement and Instrumentation Principles*. Brooks/Cole Publishing. <https://doi.org/10.1016/b978-075065081-6/50002-3>
- Arifin, W. N., & Yusoff, M. S. B. (2016). Confirmatory Factor Analysis of the Universiti Sains Malaysia Emotional Quotient Inventory Among Medical Students in Malaysia. *SAGE Open*, 6(2). <https://doi.org/10.1177/2158244016650240>
- Astuti, N. D., Hajaroh, M., Prihatni, Y., Setiawan, A., Setiawati, F. A., & Retnawati, H. (2024). Comparison of KMO Results, Eigen Value, Reliability, and Standard Error of Measurement: Original & Rescaling Through Summated Rating Scaling. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia*, 13(2), 199–217. <https://doi.org/10.15408/jp3i.v13i2.36684>

- Bean, G. J., & Bowen, N. K. (2021). Item Response Theory and Confirmatory Factor Analysis: Complementary Approaches for Scale Development. *Journal of Evidence-Based Social Work (United States)*, 18(6), 597–618. <https://doi.org/10.1080/26408066.2021.1906813>
- Beavers, A. S., Lounsbury, J. W., Richards, J. K., Huck, S. W., Skolits, G. J., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation*, 18(6), 1–13.
- Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition*, 9(2), 137–151. <https://doi.org/10.1016/j.jarmac.2019.12.009>
- Cohen R. J., & Swerdlik M. E. (2018). *Psychological Testing and Assessment: An Introduction to Tests and Measurement* (7th Edition). McGraw-Hill.
- de Jong, T., & van der Meij, J. (2012). Learning with Multiple Representations. *Encyclopedia of the Sciences of Learning, June 2016*, 2026–2029. https://doi.org/10.1007/978-1-4419-1428-6_485
- DeMars, C. (2008). Scoring Multiple Choice Items: A Comparison of IRT and Classical Polytomous and Dichotomous Methods. *Annual Meeting of the National Council of ...* <https://www.jmu.edu/assessment/CED NCME Paper 08.pdf>
- DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, 23(3), 225–241. <https://doi.org/10.1177/073428290502300303>
- Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. In *Item Response Theory for Psychologists* (1st Editio). L. Erlbaum Associates, Mahwah, N.J. <https://doi.org/10.4324/9781410605269>
- Faller, H., Kohlmann, T., Zwingmann, C., & Maurischat, C. (2006). Exploratory and confirmatory factor analysis. *Rehabilitation*, 45(4), 243–248. <https://doi.org/10.1055/s-2006-940029>
- Garcia, E., Aryal, S., Spence-Almaguer, E., Rohr, D., & Walters, S. T. (2018). Use of the IRT Model to Validate Test Items from a Technology Assisted Health Coaching Program. *Open Journal of Statistics*, 08(03), 519–532. <https://doi.org/10.4236/ojs.2018.83034>
- Ghazali, N., & Nordin, M. S. (2019). Measuring meaningful learning experience: Confirmatory factor analysis. *International Journal of Innovation, Creativity and Change*, 9(12), 283–296.
- Hair, J. F., Hult, G. T. M., Ringle, C. M., & Rstedt, M. S. (2013). A primer on partial least squares structural equation modeling (PLS-SEM). In *Sage* (Vol. 46, Issues 1–2). <https://doi.org/10.1016/j.lrp.2013.01.002>
- Hambleton, R. K., & Jones, R. W. (1993a). *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development*. 38–47.
- Hambleton, R. K., & Jones, R. W. (1993b). Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12(3), 38–47. <https://doi.org/10.1111/j.1745-3992.1993.tb00543.x>
- Hambleton, R. K., & Slater, S. C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment*, 13(1), 21–28. <https://doi.org/10.1027/1015-5759.13.1.21>
- Hambleton, R. K., Swaminathan, H., & Rogers, D. J. (1992). Fundamentals of item response theory. In *Choice Reviews Online* (Vol. 29, Issue 07). Sage Publications Inc. <https://doi.org/10.5860/choice.29-4185>

- Herwin, & Nurhayati, R. (2021). Measuring students' curiosity character using confirmatory factor analysis. *European Journal of Educational Research*, 10(2), 773–783. <https://doi.org/10.12973/EU-JER.10.2.773>
- Husain, H., & Aziz, H. (2022). Exploratory Factor Analysis (Efa) and Confirmatory Factor Analysis (Cfa) To Measure the Validity and Reliability Constructs of Historical Thinking Skills, Tpack and Application of Historical Thinking Skills. *International Journal of Education, Psychology and Counseling*, 7(46), 608–623. <https://doi.org/10.35631/ijepc.746046>
- Jiang, G., Tan, X., Wang, H., Xu, M., & Wu, X. (2023). Exploratory and confirmatory factor analyses identify three structural dimensions for measuring physical function in community-dwelling older adults. *PeerJ*, 11. <https://doi.org/10.7717/peerj.15182>
- Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, 15(1), 136–153. <https://doi.org/10.1080/10705510701758406>
- Kang, C., Huang, J., Liu, Y., & Yin, H. (2025). Development and validation of a generic self-assessment scale for K-12 teachers as feedback givers: Insights from item response theory and factor analysis. *Humanities and Social Sciences Communications*, 12(1). <https://doi.org/10.1057/s41599-025-04927-4>
- Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, 18(2), 212–228. <https://doi.org/10.1080/10705511.2011.557337>
- Kyriazos, T. A., & Stalikas, A. (2018). Applied Psychometrics: The Steps of Scale Development and Standardization Process. *Psychology*, 09(11), 2531–2560. <https://doi.org/10.4236/psych.2018.911145>
- Lauwaert, P. (2023). On Validity. *Studies in Applied Linguistics and TESOL*, 23(1), 18–36. <https://doi.org/10.52214/salt.v23i1.11804>
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13(4), 517–549. <https://doi.org/10.1177/001316445301300401>
- Lovric, M. (2011). International Encyclopedia of Statistical Science. *International Encyclopedia of Statistical Science*, March. <https://doi.org/10.1007/978-3-642-04898-2>
- Malone, K. L., Boone, W. J., Stammen, A., Schuchardt, A., Ding, L., & Sabree, Z. (2021). Construction and Evaluation of an Instrument to Measure High School Students Biological Content Knowledge. *Eurasia Journal of Mathematics, Science and Technology Education*, 17(12). <https://doi.org/10.29333/EJMSTE/11376>
- Min, S., & Aryadoust, V. (2021). A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation*, 68(December 2020), 100963. <https://doi.org/10.1016/j.stueduc.2020.100963>
- Odoi, B., Twumasi-Ankrah, S., Samita, S., & Al-Hassan, S. (2022). The Efficiency of Bartlett's Test using Different forms of Residuals for Testing Homogeneity of Variance in Single and Factorial Experiments-A Simulation Study. *Scientific African*, 17, e01323. <https://doi.org/10.1016/j.sciaf.2022.e01323>
- Ofosu, E. K., Owusu-darko, I., & Abubakar, G. A. (2020). Effect of Multiple Representation – Based Instructions (MR-BI) on SHS Students' Ability to Solve Problems on Linear Functions and Their Applications. *International Journal of Research and Scientific Innovation (IJRSI)*, 7(9), 234–239.

- Ohiri, S. C., Ihebom, D., & Nnennaya, C. (2024). Psychometric Properties of a Test: An Overview. *International Journal of Research Publication and Reviews*, 5(2), 2217–2224. <https://doi.org/10.55248/gengpi.5.0224.0539>
- Orcan, F. (2018). Exploratory and Confirmatory Factor Analysis: Which One to Use First? *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 9(4), 414–421. <https://doi.org/10.21031/epod.394323>
- Phanniphong, K., & Na-Nan, K. (2025). Development and validation of a factor analysis-validated comprehensive scale for measuring innovative work behavior. *Sustainable Futures*, 9(May 2023). <https://doi.org/10.1016/j.sftr.2025.100704>
- Plucker, J. A. (2003). Exploratory and Confirmatory Factor Analysis in Gifted Education: Examples with Self-Concept Data. *Journal for the Education of the Gifted*, 27(1), 20–35. <https://doi.org/10.1177/016235320302700103>
- Said, H., Badru, B. B., & Shahid, M. (2011). Confirmatory Factor Analysis (Cfa) for testing validity and reliability instrument in the study of education. *Australian Journal of Basic and Applied Sciences*, 5(12), 1098–1103.
- Stefanel, A. (2019). Graph in Physics Education: From Representation to Conceptual Understanding. *Mathematics in Physics Education*, 195–231. https://doi.org/10.1007/978-3-030-04627-9_9
- Svensson, K., & Campos, E. (2022). Comparison of two semiotic perspectives: How do students use representations in physics? *Physical Review Physics Education Research*, 18(2), 20120. <https://doi.org/10.1103/PhysRevPhysEducRes.18.020120>
- Taherdoost, H. (2018). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *SSRN Electronic Journal*, September. <https://doi.org/10.2139/ssrn.3205040>
- Thissen, D. (2022). Latent Variable Estimation in Factor Analysis and Item Response Theory. *Chinese/English Journal of Educational Measurement and Evaluation*, 3(3). <https://doi.org/10.59863/optz4045>
- Vorapongsathorn, T., Taejaroenkul, S., & Viwatwongkasem, C. (2004). A Comparison of Type I Error and Power of Bartlett's Test, Levene's Test and Cochran's Test under Violation of Assumptions. *Research Design & Statistics*, 26(4), 537–547.
- Walton, D. M., Nazari, G., Bobos, P., & MacDermid, J. C. (2023). Exploratory and confirmatory factor analysis of the new region-generic version of Fremantle Body Awareness—General Questionnaire. *PLoS ONE*, 18(3 March), 1–14. <https://doi.org/10.1371/journal.pone.0282957>
- Weyers, J., König, J., Santagata, R., Scheiner, T., & Kaiser, G. (2023). Measuring teacher noticing: A scoping review of standardized instruments. *Teaching and Teacher Education*, 122. <https://doi.org/10.1016/j.tate.2022.103970>
- Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, 26(3), 171–177. <https://doi.org/10.3969/j.issn.1002-0829.2014.03>
- Yoon, H. G., Kim, M., & Lee, E. A. (2021). Visual representation construction for collective reasoning in elementary science classrooms. *Education Sciences*, 11(5). <https://doi.org/10.3390/educsci11050246>

About the Contributor(s)

Godelfridus Hadung Lamanepa, is currently pursuing a doctoral degree in educational research and evaluation at Universitas Negeri Yogyakarta. The focus of this study is the field of educational measurement. Some of the research conducted is related to the field of physics education assessment. The researcher continues to conduct research in the field of evaluation, as outlined in the roadmap for the research field.

Email: godelfridushadung.2022@student.uny.ac.id

ORCID: <http://orcid.org/0009-0008-8800-209X>

Edi Istiyono, is a professor of physics education measurement at Universitas Negeri Yogyakarta. He has served as the coordinator of the Master of Education Research and Evaluation study program and is currently the coordinator of S3 Educational Research and Evaluation. Research conducted around the construction and development of physics education instruments. Various works have been published in multiple journals.

Email: edi_istiyono@uny.ac.id

ORCID: <https://orcid.org/0000-0001-6034-142X>

Raden Rosnawati, is a Lecturer in the Doctoral Program in the Department of Educational Research and Evaluation at Universitas Negeri Yogyakarta. She teaches Master's and Doctoral students in the Educational Research and Evaluation study program. In addition to having research experience, she has also written many articles on psychometrics. Her current research focuses on test development in mathematics courses.

Email: rosnawati@uny.ac.id

ORCID: <https://orcid.org/0000-0002-8841-0412>

Ayen Arsisari, is a doctoral student majoring in educational research and evaluation at the Graduate School of Universitas Negeri Yogyakarta.

Email: ayenarsisari.2022@student.uny.ac.id

ORCID: <https://orcid.org/0009-0008-6669-200X>

Fitriyani Hali, is a doctoral student majoring in educational research and evaluation at the Graduate School of Universitas Negeri Yogyakarta.

Email: fitriyanihali.2022@student.uny.ac.id

ORCID: <https://orcid.org/0009-0000-4384-0892>

Publisher's Note: *The opinions, statements, and data presented in all publications are solely those of the individual author(s) and contributors and do not reflect the views of Universitepark, EDUPIJ, and/or the editor(s). Universitepark, the Journal, and/or the editor(s) accept no responsibility for any harm or damage to persons or property arising from the use of ideas, methods, instructions, or products mentioned in the content.*
