Research Article https://doi.org/10.12973/eu-jer.12.1.371

European Journal of Educational Research

Volume 12, Issue 1, 371 - 385.

ISSN: 2165-8714

http://www.eu-jer.com/

The Development of a Four-Tier Diagnostic Test Based on Modern Test Theory in Physics Education

Edi Istiyono*
Universitas Negeri Yogyakarta, INDONESIA

Wipsar Sunu Brams Dwandaru
Universitas Negeri Yogyakarta, INDONESIA

Kharisma Fenditasari
SMA Muhammadiyah 1 Pasuruan, INDONESIA

Made Rai Suci Shanti Nurani Ayub
Universitas Kristen Satya Wacana, INDONESIA

Duden Saepuzaman
Universitas Pendidikan Indonesia, INDONESIA

Received: May 25, 2022 ▪ Revised: August 24, 2022 ▪ Accepted: December 27, 2022

Abstract: Diagnostic tests are generally two or three-tier and based on classical test theory. In this research, the Four-Tier Diagnostic Test (FTDT) was developed based on modern test theory to determine understanding of physics levels: scientific conception (SC), lack of knowledge (LK), misconception (MSC), false negatives (FN), and false positives (FP). The goals of the FTDT are to (a) find FTDT constructs, (b) test the quality of the FTDT, and (c) describe students' conceptual understanding of physics. The development process was conducted in the planning, testing, and measurement phases. The FTDT consists of four-layer multiple-choice with 100 items tested on 700 high school students in Yogyakarta. According to the partial credit models (PCM), the student's responses are in the form of eight categories of polytomous data. The results of the study show that (a) FTDT is built on the aspects of translation, interpretation, extrapolation, and explanation, with each aspect consisting of 25 items with five anchor items; (b) FTDT is valid with an Aiken's V value in the range of 0.85-0.94, and the items fit PCM with Infit Mean Square (INFIT MNSQ) of 0.77-1.30, item difficulty index of 0.12-0.38, and the reliability coefficient of Cronbach's alpha FTDT is 0.9; (c) the percentage of conceptual understanding of physics from large to small is LK type 2 (LK2), FP, LK type 1 (LK1), FN, LK type 3 (LK3), SC, LK type 4 (LK4), and MSC. The percentage sequence of MSC based on the successive material is momentum, Newton's law, particle dynamics, harmonic motion, work, and energy. In addition, failure to understand the concept sequentially is due to Newton's law, particle dynamics, work and energy, momentum, and harmonic motion.

Keywords: *Developing test, Four-Tiers Diagnostic Test, modern test theory.*

To cite this article: Istiyono, E., Dwandaru, W. S. B., Fenditasari, K., Ayub, M. R. S. S. N., & 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

Introduction

Understanding concepts in physics learning is important and fundamental (Huda et al., 2017). Evidence shows that the results of learning fundamental concepts in physics are still low (Prahani et al., 2022; Pratama & Retnawati, 2018; Saepuzaman et al., 2019). According to the overall student learning outcomes, students do not comprehend Raschthe fundamentals of physics (Rosyid et al., 2013; Suwarto, 2013). This low response from students arises when they have misconceptions or biases about the subject matter (Saepuzaman et al., 2019; Tiandho, 2018; Zukhruf et al., 2017). Misconceptions that are not overcome lead to learning difficulties in understanding the next concept. Therefore, developing tools to diagnose student concept understanding is very important.

Diagnosing students' conceptual understanding involves an instrument (Bennett, 2014; Saepuzaman & Karim, 2016). The instruments for analyzing students' conceptual understanding can be mind mapping (Kandil İngeç, 2009) or diagnostic tests (Adodo, 2013). The model of diagnostic multiple-choice questions accompanied by justifications and confidence questions is a tool that can be used to deepen students' understanding of concepts (Caleon & Subramaniam, 2010). Questions in the form of a two-tier (Tsui & Treagust, 2010), three-tier (Ratnaningdyah, 2018), and four-tier (Fariyani, 2015) diagnostic test can be used. In addition, the Four-Tier Diagnostic Test (FTDT) is expected to recognize students' concepts better than other multi-tests because the FTDT includes additional confidence questions at each level.

* **Corresponding author:**
Edi Istiyono, Universitas Negeri Yogyakarta, Jl. Colombo No.1 Karangmalang, Yogyakarta 55281, Indonesia. ✉ edi_istiyono@uny.ac.id

FTDT can minimize guessing answers (Gurel et al., 2015). In addition, the FTDT can also make a difference in student's knowledge so that the extent of students' misconceptions is known.

Diagnostic tests are tools used to recognize learning difficulties (Miller et al., 2009), and teachers use them to find students' incorrect answers. Teachers use a variety of methods to ensure students' understanding (Grigorovitch, 2014). These attempts are ineffective if the instructor does not have a comprehensive knowledge of the material's conceptualization by the students. An overview of the student's understanding of the concept will be accepted if the teacher assesses the student's learning difficulties.

The teacher assesses learning difficulties using diagnostic tests. Diagnostic tests show whether students understand a concept correctly or not. Diagnostic tests can also reveal students' misconceptions and why they have misconceptions about a scientific understanding (Gierl, 2007). Using diagnostic tests, teachers can recognize students' learning problems or difficulties. By identifying students' misunderstandings, diagnostic tests may also be used to organize subsequent attempts to correct them.

The four-tier or four-layer multiple-choice diagnostic test comprises answer choices in the first layer, as in general, in multiple-choice tests. The second layer includes the confidence level of the answers in the first layer. The third layer contains reasons, and the fourth tier is the level of belief in reasons. A four-tier multiple-choice diagnostic test can be the expansion of a three-tier diagnostic test by adding a level of confidence to each answer and reason. The third stage covers pertinent concepts that authenticate the response in the first stage, but the fourth stage delivers assurance about the replies in the third stage.

The categorization of FTDT used will refer to Gurel et al. (2015). It developed in categorization by paying attention to the weighting of answers and reasons made by Istiyono et al. (2014). So it is obtained to categorize the misconceptions of the four-tier test according to Gurel et al. and Istiyono et al. as in Table 1.

Table 1. Four-Tier Test Decision According to Gurel et al. (2015) and Istiyono et al. (2014)

Answer	Level of Confidence	Reason	Level of Confidence	Decision
Correct	Sure	Correct	Sure	Scientific Conception (SC)
Correct	Sure	Correct	Not sure	Lack of Knowledge type 1(LK1)
Correct	Not sure	Correct	Sure	Lack of Knowledge type 1(LK1)
Correct	Not sure	Correct	Not sure	Lack of Knowledge type 1(LK1)
Correct	Sure	Wrong	Not sure	Lack of Knowledge type 2 (LK2)
Correct	Not sure	Wrong	Sure	Lack of Knowledge type 2 (LK2)
Correct	Not sure	Wrong	Not sure	Lack of Knowledge type 2 (LK2)
Wrong	Sure	Correct	Not sure	Lack of Knowledge type 3 (LK3)
Wrong	Not sure	Correct	Sure	Lack of Knowledge type 3 (LK3)
Wrong	Not sure	Correct	Not sure	Lack of Knowledge type 3 (LK3)
Wrong	Sure	Wrong	Not sure	Lack of Knowledge type 4 (LK4)
Wrong	Not sure	Wrong	Sure	Lack of Knowledge type 4 (LK4)
Wrong	Not sure	Wrong	Not sure	Lack of Knowledge type 4 (LK4)
Wrong	Sure	Wrong	Sure	Misconception (MSC)
Wrong	Sure	Correct	Sure	False Negative (FN)
Correct	Sure	Wrong	Sure	False Positive (FP)

The use of modern test theory is the proper explanation to obtain valid and reliable test development (Krathwohl, 2002) because the analysis using modern test theory emphasizes more on items and item responses (Lane et al., 2016). The use of modern test theory using item response theory (IRT) characterizes latent trait theory or item characteristic curve (ICC). The use of IRT aims to overcome the weaknesses found in classical measurements. Opportunities answered correctly $P(\theta)$ an item parameter: in the form of discrimination power (a), item difficulty index (b), and guessing (c), as well as the characteristics or parameters of the test taker, always correlated with a model formula that must be followed by grouping test items or groups of test takers (Hambleton et al., 1991).

The partial credit model (PCM) is a polytomous scoring model used to evaluate item response models with more than two categories (Retnawati, 2014). The PCM was developed from the Rasch model (RM). The RM is used for dichotomous score data. Inline dichotomous score data, PCM was considered for categorical or polytomous score data. Scoring of categories corresponding to the scoring FTDT.

The pandemic condition caused by the COVID-19 virus also disturbs the learning assessment system. Face-to-face learning assessment is not possible in this situation. Therefore, an online-based assessment model is needed so that learning continues.

Based on the explanation, it is very important to investigate the development of the FTDT online paper-based test (FTDT-O_PBT) in an effective and targeted Google form using modern theory tests in its examination. The development of FTDT-O_PBT was carried out to produce the correct test construct, good quality instrument, accurate measurement results, and effectively identify students' conceptual understanding appropriately. This study aims to develop a FTDT based on physics learning using modern test theory analysis.

Methodology

Research Design

The development of FTDT online instrument was developed in an FTDT-O_PBT. The validity and quality of the contents of the instrument will be tested by expert judgments, seven experts consisting of material experts, measurement experts, and practitioners of physics education. The Orlando-Antonio test development model used is the instrument development method, which was improved into three steps, (a) test planning, (b) test trials, and (c) measurement and interpretation.

The first stage was the test planning stage or the question construction stage. The first step was formulating the matrix and test grid, item writing, and item validation. In the next step, before being tested, the instrument was passed content validation by experts (expert judgment), seven experts consisting of material experts, measurement experts, and physics education practitioners.

Sample and Data Collection

The FTDT-O_PBT test instrument was given to 700 students of class X high school students in Yogyakarta province at 15 schools. The scoring rubric for FTDT-O_PBT questions used eight categories referring to the Four-tier test decision in Table 2 (Istiyono et al., 2014).

Table 2. FTDT scoring

Category of Concept Understanding	Scoring
Scientific Correct (SC)	7
Lack Knowledge type 1 (LK1)	6
False Positive (FP)	5
Lack Knowledge type 2 (LK2)	4
False Negative (FN)	3
Lack Knowledge type 3 (LK3)	2
Misconception (MSC)	1
Lack Knowledge type 4 (LK4)	0

The selection of school types was based on high school National Examination scores in physics with high, medium, and low rankings. The next process was the validation, measurement, and data interpretation of students' understanding concept

Analyzing of Data

The instrument's content validity, tested by expert judgments, will be examined using the Aiken equation on eq.1 (Aiken, 1985; Azwar, 2012). Aiken's validity coefficient is calculated with a score of 7 experts. Aiken's coefficient value (Aiken's V) will range from -1 to 1 (Putranta & Supahar, 2019). The coefficient value of V items is valid if it has an index V > 0.8.

$$V = \frac{\sum(r_i - l_0)}{[n(c-l)]} \; [1]$$

With r_i is the number given by the expert, l_0 is the number of the lowest validity, c is the number of the highest validity, and n is the number of experts and practitioners who carry out the assessment. The concept of content validity proposed by Aiken (1985) is influenced by the number of raters and the rating scale used.

The quality of other instruments and construct validity was also examined based on students' responses to find out whether the seven sub-aspects of formation and items have been made fit to the latent conceptual understanding by referring to the results of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The instrument's suitability was evaluated using the Rasch model from the average INFIT MNSQ (Mean of Square) value. Items fit the model if they have an INFIT MNSQ value of 0.77 to 1.33 (Mayers et al., 2002). In addition, empirical validity was demonstrated by evaluating the respondent's responses to FTDT-O_PBT. Empirical validity (Apino & Retnawati, 2016; Sumintono & Widhiarso, 2015) was determined by using IRT, in this case, is PCM analysis which the R program was used during the examination.

PCM is a development of the Rasch model of dichotomous items applied to polytomy items. According to Muraki and Bock (1997, as cited in Retnawati, 2014), the general form of PCM is as in eq.2.

$$P_{jk}(\theta) = \frac{\exp \sum_{v=0}^{k}(\theta - b_{jv})}{\sum_{h=0}^{m} exp \sum_{v=0}^{k}(\theta - b_{jv})}\ , k = 0,1,2, \dots, m\ [2]$$

With $P_{jk}(\theta)$ is the probability of participants with the ability θ gain category score k_m in item j, θ is the ability of participants, m + 1 is the number of items j, and b_{jk} is the difficulty index category k items to j.

The FTDT-O_PBT items are valid if the INFIT MNSQ value is 0.77 to 1.30 (Subali & Suyata, 2012). The level of instrument reliability is determined using the Alpha coefficient. The alpha coefficient can be used as long as each hemisphere is the same length or contains the same number of items (Azwar, 2020). The alpha coefficient can be determined using eq.3 (Azwar, 2019). The alpha reliability is in the range of values between 0 and 1. With $\rho_{xx'}$ is the coefficient reliability, α is the alpha coefficient, k is the number of gains, $\sigma_x^{x^2}$ is a variant of the test score, is a variant of the test hemisphere score with i is 1, 2, 3, and so on.

$$\rho_{xx'} \geq \alpha = \frac{\left[\frac{k}{k-1}\right]}{\sigma_0}\left[\sigma_x^2 - \sum \sigma_r^2\right]\ [4]$$

The difficulty index for each item is calculated using the following equation (5) (Baker et al., 2001):

$$P = \frac{N_p}{N}\ [5]$$

P is the *proportion* of the item difficulty index, N_p is the number of participants who answered correctly, and N is the total number of participants. Reliability for rating results from observations of several raters (Shrout & Fleiss, 1979). Observed in the ICC (Interclass Correlation Coefficient), estimated using eq.6 (Mardapi, 2008):

$$ICC = \frac{\sigma_S^2}{\sigma_S^2 + \sigma_0^2 + \sigma_e^2}\ [6]$$

σ_s the size of the study object variants, σ_0 is an instrument variant, and $\sigma_e$the Variance is due to random factors. ICC reliability will be used to estimate the reliability of the observation sheet instrument.

Items Information Function (IIF) is a method that can explain the power of an item. The information function will state how strong the contribution of the item being observed is in revealing the latent strength of the item trait and selecting the item strength of the test. The information function is expressed in eq.7 (Baker et al., 2001; Lissitz & Samuelsen, 2007)

$$I_i(\theta) = \frac{[P_i{}'(\theta)]}{P_i(\theta)Q_i(\theta)}\ [7]$$

with i is 1,2,3,..., n, $I_i(\theta)$ is a function of grain to I, $P_i(\theta)$ is the probability of participants with the ability θ answering correctly item i, $P_i{}'(\theta)$ is the derivative of the function $P_i(\theta)$ to θ, and $Q_i(\theta)$is the probability of participants with the ability θ answering incorrectly item i. Test information can be written mathematically as in eq.8,

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)\ [8]$$

Findings / Results

The results include the construct results, FTDT quality, and a description of the understanding of the concept. These results are the development of questions that have been tested.

FTDT Construct

The construct results found that the FTDT-PBT test instrument developed consisted of 100 test items, 5 of which were anchor items. FTDT-PBT test was developed based on aspects of translation, interpretation, extrapolation, and explanation, as stated in Table 3.

Table 3. The FTDT-PBT test matrix

Aspect	Sub aspect	Indicator	Materials					
			1	2	3	4	5	Total
Translation	Translating	Translating physical symptoms	1A, 11B, 6C, 11D	2A, 15B, 18C, 1D	3A*, 17B*, 20C*, 24D*	4A, 13B, 19C, 14D	5A, 5B, 3C, 13D	20
	Interpreting	Interpreting irrelevant information	6A*, 6B*, 11C*, 8D*	7A, 1B, 1C, 15D	8A, 12B, 4C, 9D	9A, 23B, 16C, 5D	10A, 14B, 22C, 3D	20
Interpretation	Interpreting	Interpreting values based on data	11A, 16B,24C, 19D	12A, 19B,21C, 18D	13A, 4B, 12C, 2D	14A*, 2B*, 23C*, 23D*	15A, 9B, 8C, 17D	20
Extrapolation	Extrapolating	Extrapolating relationships between variables	16A, 20B, 2C, 4D	17A*, 8B*, 9C*, 6D*	18A, 10B, 7C, 20D	19A, 7B, 13C, 16D	20A, 18B, 14C, 10D	20
Explaining	Explaining	Explaining problem solving in detail	21A, 3B, 17C, 22D	22A, 21B, 15C, 21D	23A, 22B, 25C, 12D	24A, 25B, 5C, 7D	25A*, 24B*, 10C*, 25D*	20
Total								100

Description: * = item question anchor 1= Particle Dynamics, 2= Newton's Laws of Gravity, and Kepler's Laws, 3=Concepts of Work and Energy, 4= Momentum and Impulse, 5= Harmonic Vibration

FTDT Quality

The FTDT quality test results for Aiken's V score are shown in Table 4; the total variance extraction value in Table 5; the factor loading value in Table 6; the estimated item and FTDT test scores in PCM 1- PL in Table 7; the reliability value of the FTDF question in Table 8; and the item difficulty index in Table 9 below.

Table 4. The value of Aiken's V items in each package

Package	Minimum Value	Maximum Value	Average of Aiken's V	Result
A	.848	.935	.902	Valid
B	.855	.942	.894	Valid
C	.870	.949	.923	Valid
D	.841	.928	.894	Valid

Table 5. The total variance extraction sum of squared loading

Component	Package A		Package B		Package C		Package D	
	Total	% of Variance (%)	Total	% of Variance (%)	Total	% of Variance (%)	Total	% of Variance (%)
1	**7.809**	**31.235**	**7.268**	**29.072**	**7.673**	**30.693**	**8.581**	**34.326**
2	**1.538**	**6.153**	**1.357**	**5.427**	**1.636**	**6.546**	**1.295**	**5.181**
3	1.248	5.992	1.112	5.448	1.096	5.384	1.150	5.599
4	1.093	5.372	1.046	5.182	1.077	5.309	1.062	5.248
5	1.024	5.097	1.011	5.042	1.016	5.063	1.011	5.045

Table 6. The RMSEA fit index

RMSEA fit index	Value	Standard value	Description
Package A	0.057	≤ 0.08	Fit
Package B	0.036	≤ 0.08	Fit
Package C	0.059	≤ 0.08	Fit
Package D	0.039	≤ 0.08	Fit

Table 7. Item estimation results and FTDT-PBT physics using PCM 1-PL

Package A			**Package B**			**Package C**			**Package D**		
	Aspect	**Value**		**Aspect**	**Value**		**Aspect**	**Value**		**Aspect**	**Value**
Outfit MNSQ	Person	1.101	Outfit MNSQ	Person	0.982	Outfit MNSQ	Person	1.111	Outfit MNSQ	Person	1.211
	Item	1.112		Item	0.970		Item	1.121		Item	1.132
Outfit t	Person	-1.110	Outfit t	Person	0.846	Outfit t	Person	1.110	Outfit t	Person	-1.091
	Item	-1.140		Item	0.987		Item	1.104		Item	-1.062
INFIT MNSQ	Person	0.90	INFIT MNSQ	Person	1.087	INFIT MNSQ	Person	1.071	INFIT MNSQ	Person	1.072
	Item	0.98		Item	1.081		Item	1.082		Item	1.181

Table 8. The FTDT-O_PBT reliability test

Test package	**Cronbach's Alpha**
A	.907
B	.897
C	.905
D	.920

Table 9. Range of item difficulty index for each package

	Package A	**Package B**	**Package C**	**Package D**
Minimum	-0 .12	-0 .09	-0 .11	- 0.10
Maximum	.38	.38	.38	.38

The Description of Understanding the Concept of Physics

The result of understanding the physics concept based on student responses is shown in Figure 1. Simultaneously, the distribution of conceptual understanding of physics for physics material can be seen in Table 10.

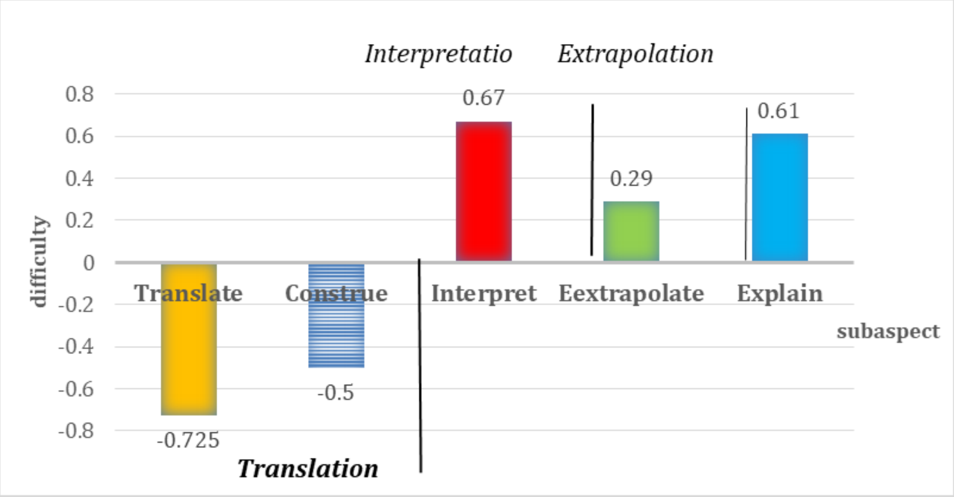


Figure 1. The Difficulty Index in Understanding Concepts

Table 10. The distribution of conceptual understanding of physics class in percentage

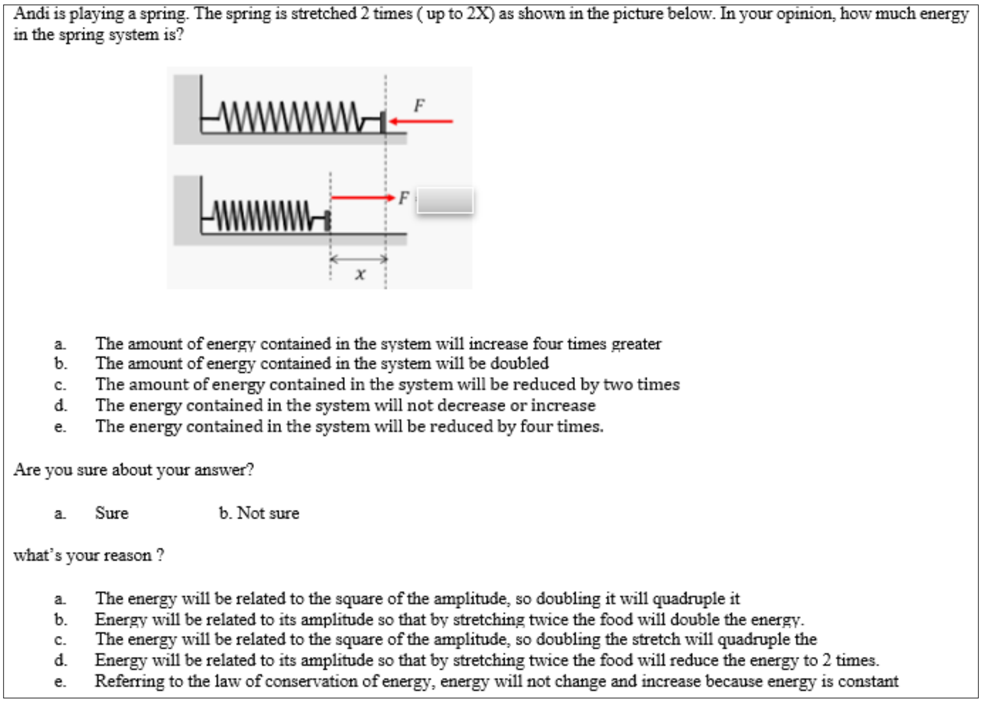
	SC (%)	**LK1 (%)**	**FP (%)**	**LK2 (%)**	**FN (%)**	**LK3 (%)**	**MSC (%)**	**LK4 (%)**
Particle dynamics	12.41	12.49	12.50	13.60	12.97	12.43	11.90	11.70
Newton's law	10.89	12.19	13.76	14.11	13.81	11.63	12.06	11.55
Work and Energy	12.24	13.23	13.02	13.30	13.15	12.51	11.72	10.83
Impulse Momentum	13.15	13.89	12.46	12.62	11.85	12.68	12.76	10.59
Harmonic Motion	12.21	13.14	14.06	13.05	12.28	11.78	12.29	11.19

Discussion

The FTDT Construct

The synthesis of constructs arranged in Table 3 refers to Bloom's (Crumb, 1983) conceptual understanding theory and its revision (Anderson & Krathwohl, 2010) with aspects of translation, interpretation, extrapolation, and explanation. As a hierarchical structure, Bloom's taxonomy can identify student abilities (Tiandho, 2018) from low to high levels (Yanuike et al., 2017). The students' conceptual understanding ability, which is measured, includes sub-aspects (a) translating with the success indicators of students being able to translate physical symptoms, and (b) interpreting with students' success indicators, namely students being able to interpret irrelevant information. Aspects of interpretation will include aspects of interpreting with indicators of success; students can interpret the value of physics based on existing data. The extrapolation aspect indicates students' ability to extrapolate the relationships between variables. The last aspect is an explanation with sub-aspects of solving the problem in detail with indicators that students can explain the solution. These aspects result from synthesizing students' conceptual understanding abilities that often appear in students (Holme, 2015). Each aspect is built with a 25-item test with five anchor items.

FTDT consists of the dynamics of particles, Newton's law of gravity, Kepler's law, the concepts of work and energy, momentum and impulse, and harmonic vibrations material. The questions are made in four packages, A, B, C, and D, with the same test achievement indicators between packages. The questions are made multiple-choice with a four-tier test, with the FTDT item model shown in Figure 2. The development of this FTDT question is appropriate for diagnosing students' conceptual understanding by adding confidence questions at each level, both in answers and reasons. It minimizes the guessing responses (Gurel et al., 2015). Explanations are made based on the findings of the students' incorrect answers.



Andi is playing a spring. The spring is stretched 2 times (up to 2X) as shown in the picture below. In your opinion, how much energy in the spring system is?

a. The amount of energy contained in the system will increase four times greater
b. The amount of energy contained in the system will be doubled
c. The amount of energy contained in the system will be reduced by two times
d. The energy contained in the system will not decrease or increase
e. The energy contained in the system will be reduced by four times.

Are you sure about your answer?

a. Sure b. Not sure

what's your reason ?

a. The energy will be related to the square of the amplitude, so doubling it will quadruple it
b. Energy will be related to its amplitude so that by stretching twice the food will double the energy.
c. The energy will be related to the square of the amplitude, so doubling the stretch will quadruple the
d. Energy will be related to its amplitude so that by stretching twice the food will reduce the energy to 2 times.
e. Referring to the law of conservation of energy, energy will not change and increase because energy is constant

Figure 2. Examples of FTDT Items

The Quality of FTDT

The results of the validator assessment in Table 3 were analyzed with the validity of Aiken's V in Equation 1, with results as shown in Table 4. It can be seen in table 4 that the analysis results can be categorized as valid if it meets Aiken's V coefficient limit for 4 rating scales and seven raters are 0, 76. Table 4 shows that the range of Aiken's V value of each item in packages A, B, C, and D is > 0,8, so it can be concluded that the FTDT items in each package O_PBT are valid and usable.

Factor extraction results for four packages using EFA can be seen in Table 5. The seven established sub-aspects produce five forming factors with eigenvalues > 1 for each package, as shown in Table 5, so that the number of components with eigenvalues > 1 will be intended as the number of extracted factors that are then allocated to the model (Bauldry, 2015) on the FTDT-O_PBT instrument. Table 5 shows that each package produces five factors with a total eigenvalue of >1. Fabrigar & Wegener (Bauldry, 2015) state that the number of components with eigenvalues > 1 with a variance percentage > 5% will be intended as the number of extracted factors allocated to the model.

Confirmatory Factor Analysis (CFA) results for packages A, B, C, and D of the five factors with 25 items. The loading factor value on each package with standardized loading estimate ≥ .50) and $_{count}$ (> 1.96) on each package, resulting p-value greater than α with the RMSEA package, as shown in Table 6. Convergent Validity analysis is carried out to determine that the items from a latent construct have been assembled with a high proportion of variants (Ghozali & Fuad, 2005).

The fitness of the model items on the FTDT-O_PBT test, which was analyzed using the R Program, showed that all items in Packages A, B, C, and D fit the model. The summary of the estimation results can be seen in Table 7. The goodness of fit results can be seen in the INFIT MNSQ and OUTFIT MNSQ sections (Subali & Suyata, 2012). Table 7 shows the compatibility of each item with PCM with INFIT MNSQ acceptance limits for all packages in the range of .77-1.30 and MNSQ OUTFIT acceptance limits in the range of .5-1.5 and t OUTFIT for all packages ≤ 2,0.

The FTDT-O_PBT instrument developed on all packages was stated to be reliably assessed from Cronbach's alpha value. The reliability values of all packages are shown in Table 8. Table 8 shows the Cronbach's alpha value (Azwar, 2020) for packages A, C, and D > from .9, which indicates the reliability criteria for the package are excellent. While the Cronbach's Alpha value of package B is.81 to.9, the item criteria are still in the outstanding category.

Based on Table 9, the difficulty index of the test items of the instrument in package A is in the range of -0.12 to 0.38. Package B is in the range of -0.09 to 0.38, package C is in the range of -0.11 to 0.38, and package D is in the range of -0.10 to 0.38. A positive sign indicates that the item is categorized as complex, and a negative sign indicates that the item is classified as easy (Baker et al., 2001). The FTDT-O_PBT items are good because they are from -2.0 to 2.0. From the explanation above, it can be concluded that FTDT-O_PBT has an item difficulty index of -0.12 to 0.38. The difficulty index of an item is a description of the item's difficulty as a whole (Azwar, 2020). The difficulty index of the FTDT-O_PBT items already represents the student's ability to answer these items in the test analysis.

The interclass correlation coefficient (ICC) observations (eq.5) of the FTDT-O_PBT test results are shown in Figures 1 and 2 using package C item 1 as an example. Figures 1 and 2 are item characteristic curve (ICC) measurements). The ICC for Item 1 of Question Package C is shown in Figure 2. The ICC for Item 1 of Package B is shown in Figure 3, indicating that respondents' abilities are based on the category of student responses (Retnawati, 2016a; Sumintono & Widhiarso, 2015).

There are eight categories for student responses from 0 to 7. As can be seen in Figure 2 and Figure 3, very low-ability students are students who score 0 and 1. In contrast, low ability students score 2 and 3, high ability students score 4 and 5, and very high ability students score 6 and 7. It shows that the higher the probability, the higher the chance of answering correctly (Crocker, 2012; Retnawati, 2016a, 2016b).

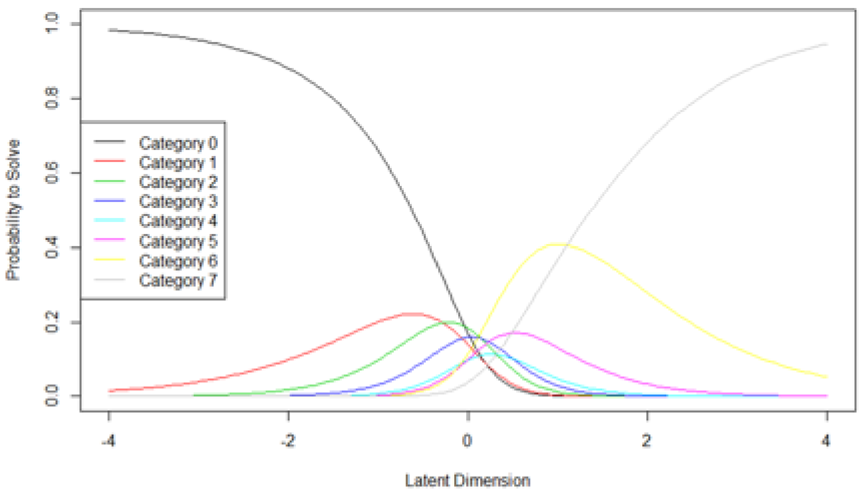


Figure 3. Characteristic Curve Item 1 Package C

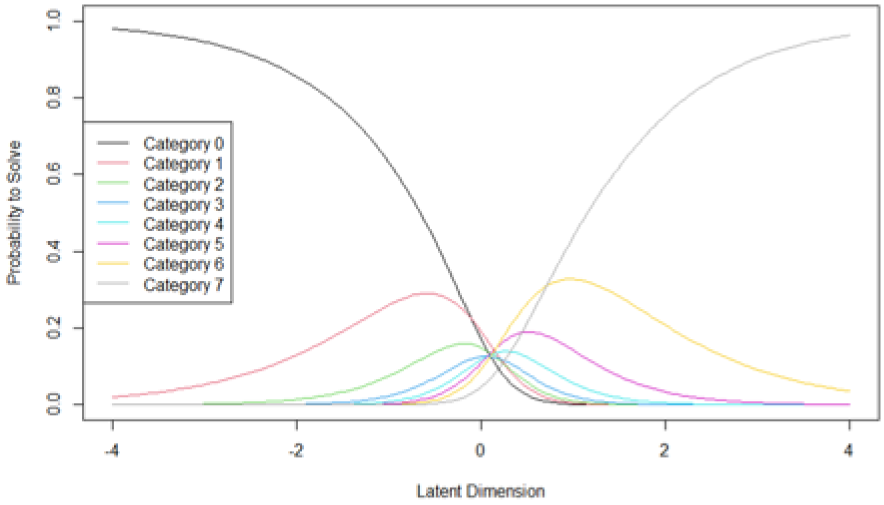


Figure 4. Characteristic Curve Item 1 Package B

The score curve in Figure 2 has a non-sequential capability category result. It shows the characteristics of item 1 in package C, but it cannot deliver good scoring. Because the higher the score, the higher the student's ability (Shrout & Fleiss, 1979). It looks different from the results in Figure 4, with the ability categories sequenced according to the student's abilities.

The result of the information value of each package is relatively similar. Each item package has a difficulty index of -2 to 2. This result shows that the developed questions can provide good information for respondents with abilities of -2 to 2 (Mardapi, 2017). The test information function image (eq.7) for packages A, B, C, and D are shown in Figure 5, Figure 6, Figure 7, and Figure 8.

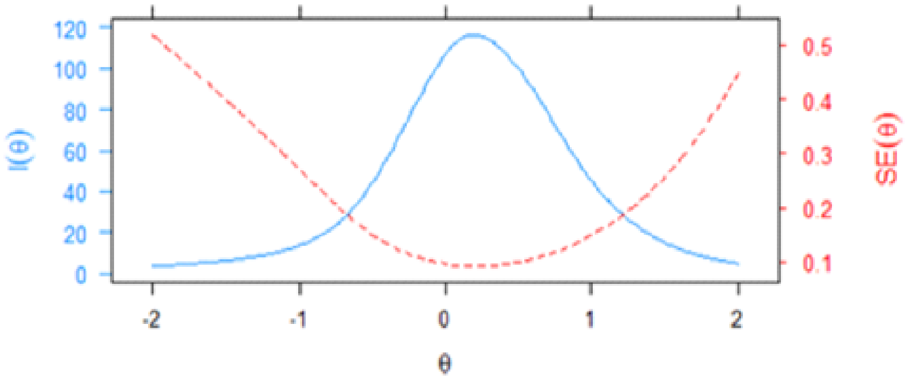


Figure 5. Information Function Plot Package A

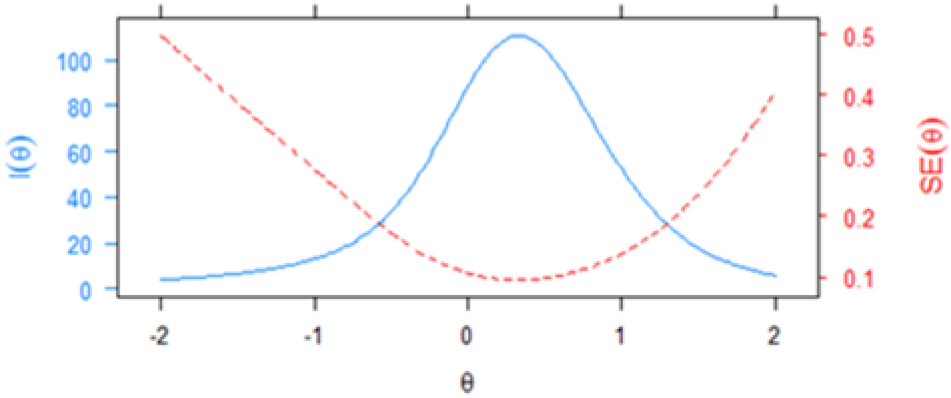


Figure 6. Information Function Plot Package B

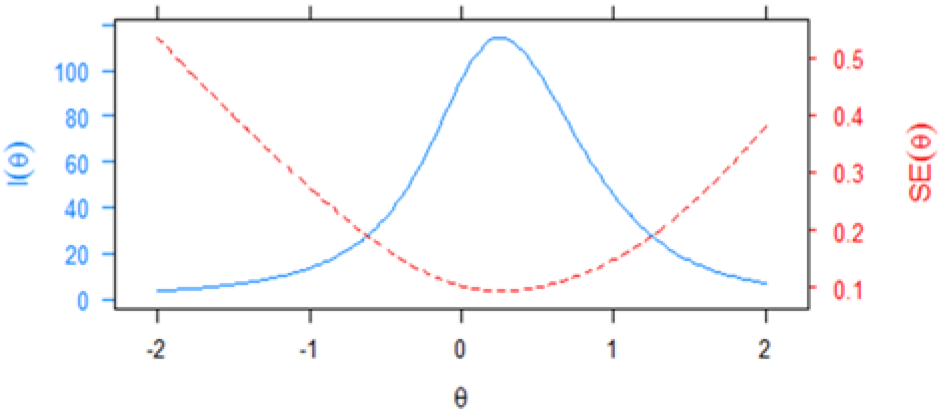


Figure 7. Information Function Plot Package C

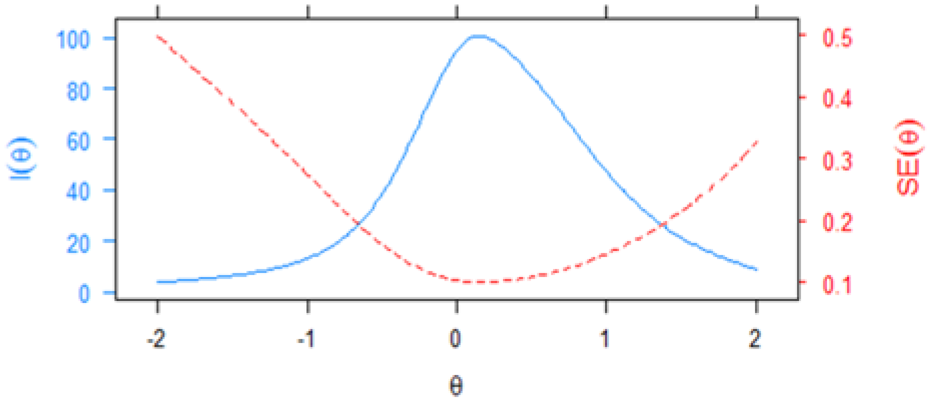


Figure 8. Information Function Plot in Package D

The respondent's score reflects the respondent's ability (Fariyani, 2015; Ratnaningdyah, 2018). So the respondent's score and ability are the respondents' parameters. Respondents' ability is a continuum from low to high (Krathwohl, 2002; Lane et al., 2016). High respondent scores indicate high ability, and low scores indicate low ability.

Description of FTDT

Overall, the respondent's data found that the difficulty index of the FTDT O_PBT translation aspect is -0,725, the interpretation aspect is 0,67, the extrapolation aspect is 0,29, and for the explaining aspect, it is 0,61. It is displayed in graphical form, as in the figure. Figure 1 presents the percentage of item difficulty index from each sub-aspect from the easiest to the most consecutively (Zulfikar et al., 2019) in the interpreting sub-aspect of explaining, extrapolating, and translating. It means that the ability to explain and translate is still relatively low (Crumb, 1983), while the ability to interpret, estimate and explain has a high score.

Based on Table 9, conclusions were drawn on the spread of the student's conceptual understanding of physics for particle dynamics materials, Newton's laws of Gravity and Keppler's law, the concepts of work and energy, momentum and impulse, and harmonic motion on the FTDT-O_PBT instrument. Table 9 shows MSC based on material percentages from the largest to the smallest: momentum-impulse, Newton's law, particle dynamics, harmonic motion, and work and energy. Misconception about the physics concept, a combination of understanding LK3, LK4, and FN obtained successively is Newton's law of particle dynamics of 37.1% work and energy of 36.85%, momentum-impulse of 34.85%, and harmonic motion of 35.25%. Enough concept understanding is the sum of the abilities of LK1, LK2 and FP obtained, respectively, Newton's law of 40.06%, the harmonic motion of 40.05%, work, and energy of 39.56%, impulse-momentum of 38,75%, and particle dynamics of 38,59%. Meanwhile, from the largest to the smallest, the Scientific Correct (SC) categories are impulse-momentum at 13.15%, particle dynamics at 12.41%, work and energy at 12.24%, and harmonic motion at 12.21%, and Newton's laws at 10.89%.

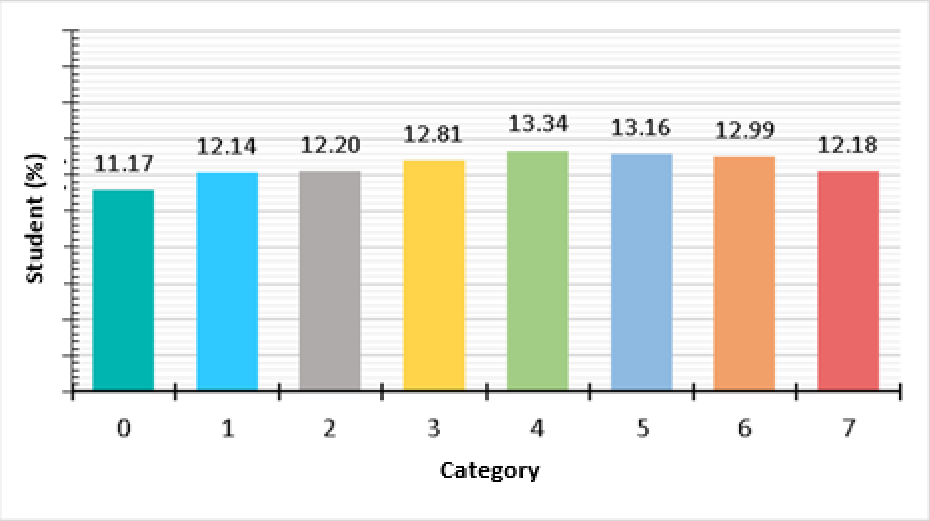


Figure 9. Mapping of Conceptual Understanding of Physics Using FTDT-PBT

The analysis and follow-up for each categorization are described below. Conceptual understanding of physics with the four-tier test scoring model shows the average categorization results as shown in the graph in Figure 8. Figure 8 shows the results of the conceptual understanding of Physics categorization with the four-tier test scoring model for SC of 12.18%, LK1 of 12.99%, FP 13.16%, LK2 13.34%, FN 12.81%, LK 12.20%, MSC 12.14%, and LK4 11.17%. Figure 8 shows that the misconception ability is the sum of the FN, LK3, and LK4, 36.18%. The ability to understand the concept, which is the sum of LK1, LK2, and FP, is 38.99%. Scientific Correct ability (SC) is 12.18%.

Students in the Scientific Correct (SC) category have correctly understood the concept (Cetin-Dindar & Geban, 2011). Students in this category have no difficulty understanding concepts to continue to the following material. Students with the category of misconceptions (MSC) may experience several possible misconceptions that require more in-response and deeper research. The misconceptions caused by associative thinking will be handled differently with students who experience misconceptions with humanistic thinking or students with wrong reasoning and intuition thinking. Students who have enough understanding of concepts understand the concepts of LK1, LK2, and FP. Students of this type have almost complete knowledge or nearly meet the understanding of scientific concepts, and only their confidence level is low. Scientific knowledge will be formed if it is believed to be accurate by students (Roberts et al., 2006; Yang et al., 2015). The category of students with correct answers and the right reason but not confident with their answers should be helped to trust their choices by encouraging self-confidence in students. Students with misconceptions conceptualize or include LK 4, LK 3, and FN categories. These students do not understand concepts (true lack of knowledge), so it is advisable to re-learn concepts. Students who experience misconceptions (MSC) respond with wrong answers and reasons but believe that such students need to be explained the correct concept.

This study also suggests that the diagnostic test used is able to identify the level of students' conceptual understanding. This can be seen from the netting of all categories of misconceptions that occur in students. So, it can be concluded overall that students still do not understand the concept of class XI physics concept. The results of this study can help in exploring concepts that have not been understood by students. This is in accordance with the research conducted by Silung et al. (2016) that diagnostic tests are able to evaluate students' misconceptions by looking at the answers, reasons and level of confidence in answering questions. In contrast to the research of Silung which used the Three-Tier Diagnostic, this study used the FTDT. The development of the four-tiers is expected to be more sensitive and responsive in identifying students' understanding of concepts. So that various conditions of understanding students' concepts can be known, in order to plan appropriate actions and treatments in an effort to improve learning outcomes.

Conclusion

Three conclusions can be drawn based on this study's data analysis and discussion. FTDT-O_PBT was developed based on the aspects sub-aspects, and materials of physics, including the translation aspects with the sub-aspects of translation and interpretation. The interpretation aspects with the sub-aspects of interpreting, the extrapolation aspects with the sub-aspects of extrapolating, and the explanation aspects with the sub-aspects of explaining. The test was developed on Newton's law, particle dynamics, work and energy, momentum, and harmonic motion. The FTDT-O_PBT consists of 5 sets of 25 items and five anchor items for each set. Second, FTDT-O_PBT has proven content validity with an Aiken's V of 0.848-0.935, the items fit PCM with an INFIT MNSQ of 0.77-1.30, and the item difficulty index ranges from -0.12 to 0.38, and the Cronbach Alfa reliability coefficient is 0.9, indicating that the FTDT-O_PBT is valid and reliable. The third one, the FTDT-O_PBT instrument, can describe the conceptual understanding of physics of high school students; which are SC (2.18%), LK1 (12.99%), FP (13.16%), LK2 (13.34%), FN (12.81%), LK3 (12.20%), MSC (12.14%), LK4 (11.17%), with the largest to the smallest number being LK2, FP, LK1, FN, LK3, SC, MSC, and LK4. Based on the material, misconceptions (MSC) are from largest to smallest number momentum-momentum, followed by Newton's law, particle dynamics, work, and energy. Conceptual misconceptions (FN, LK3 dan LK4) of 36.18% of students are related to Newton's law, particle dynamics, work and energy, momentum, and harmonic motion. It is enough if 38.99% of students understand the concepts (LK1, LK2 dan FP), from the largest to the smallest; Newton's law, harmonic motion, work and energy, momentum, and particle dynamics. Scientific concepts (SC) are 12.18% of students with material momentum-momentum, particle dynamics, work and energy, harmonic motion, and Newton's law.

Recommendations

These findings can be used as a starting point in developing a stratified multiple-choice test instrument. It is very important to diagnose students' conceptual difficulties as early as possible. Further research is expected to examine the appropriate treatment or learning for each group of students' conceptions, ranging from misconceptions to the type of Lack of Knowledge.

This expansion of the category of scientific concepts allows for more precise measurements. The expansion of the material allows it to be developed in other materials and lessons. In addition, it can also be used to diagnose students' scientific concept abilities, both material, and student categories.

Limitations

Instrument analysis was performed using PCM analysis. The item parameters analyzed are only the difficulty level, while the discriminative power parameters were not examined. To obtain more certainty regarding the obtained results, we performed an analysis with the GPCM analysis to determine other parameters for each item, not only the difficulty level but also the discriminative power of the items. Thus, the quality of the items in terms of their parameters becomes more valid as an instrument component. In addition, this instrument was tested on students in Yogyakarta for specific reasons that might differ from the abilities of other students in Indonesia. This study only used high school students in Yogyakarta as respondents. Further research is expected to use representative high school students representing all regions in Indonesia. To represent the abilities of all Indonesian high school students so that the instrument quality would be more valid when used for students in Indonesia.

Acknowledgments

This research was carried out with financial support from the Ministry of Research, Technology, and Higher Education for 2021 Number 134/E4.1/AK.04.PT/2021.

Funding

Ministry of Research, Technology, and Higher Education for 2021.

Authorship Contribution Statement

Istiyono: Concept and design, supervision, final approval. Dwandaru: Data analysis/interpretation. Fenditasari: Collecting data and instrument. Ayub: Drafting manuscript, critical revision of the manuscript. Saepuzaman: Statistical analysis.

References

Adodo, S. O. (2013). Effects of two-tier multiple choice diagnostic assessment items on students' learning outcome in basic science technology (BST). *Academic Journal of Interdisciplinary Studies*, 2(2), 201. https://doi.org/10.5901/ajis.2013.v2n2p20

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Anderson, L. W., & Krathwohl, D. R. (2010). *Kerangka landasan untuk pembelajaran, pengajaran, dan asesmen* [Foundational framework for learning, teaching, and assessment]. Pustaka Pelajar.

Apino, E., & Retnawati, H. (2016). Creative problem solving to improve students' higher order thinking skills in mathematics instructions. In W. Warsono (Ed.), *Proceeding of 3rd International Conference on Research, Implementation and Education of Mathematics and Science* (pp. 339-346). Universitas Negeri Yogyakarta.

Azwar, S. (2012). *Reliabilitas dan validitas* [Reliability and validity] (1st ed.). Pustaka Pelajar.

Azwar, S. (2019). *Konstruksi tes kemampuan kognitif* [Cognitive ability test construction]. Pustaka Pelajar.

Azwar, S. (2020). *Reliabilitas dan validitas* [Reliability and validity] (4th ed.). Pustaka Pelajar.

Baker, M., Rudd, R., & Pomeroy, C. (2001). Relationships between critical and creative thinking. *Journal of Southern*

Agricultural Education Research, *51*(1), 173–188.

Bauldry, S. (2015). Structural equation modeling. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 615–620). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.44055-9

Bennett, D. M. (2014). The U.N. convention on the rights of persons with disabilities and U.K. mental health legislation. *British Journal of Psychiatry*, *205*(1), 76–77. https://doi.org/10.1192/bjp.205.1.76a

Caleon, I. S., & Subramaniam, R. (2010). Do students know What they know and what they don't know? Using a four-tier diagnostic test to assess the nature of students' alternative conceptions. *Research in Science Education*, *40*(3), 313–337. https://doi.org/10.1007/s11165-009-9122-4

Cetin-Dindar, A., & Geban, O. (2011). Development of a three-tier test to assess high school students' understanding of acids and bases. *Procedia-Social and Behavioral Sciences*, *15*, 600–604. https://doi.org/10.1016/j.sbspro.2011.03.147

Crocker, L. (2012). Introduction to measurement theory. In J. Green, G. Camilli & P. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 371–384). Routledge. https://doi.org/10.4324/9780203874769

Crumb, L. N. (1983). The classification of biographical dictionaries in reference collections using the library of congress classification system. *Cataloging & Classification Quarterly*, *3*(1), 41-44. https://doi.org/10.1300/J104v03n01_03

Fariyani, Q. (2015). *Pengembangan four-tier diagnostic test untuk mengungkap miskonsepsi fisika siswa sma kelas X* [Development of four-tier diagnostic test to reveal the physics misconceptions of class X high school students]. *Journal of Innovative Science Education*, *4*(2), 41–49. https://l24.im/oACm4

Ghozali, I., & Fuad, H.. (2005). *Structural equation modeling: Teori, konsep, dan aplikasi dengan program LISREL 8.54* [Structural equation modeling: Theory, concepts, and applications with programs LISREL 8.54]. Badan Penerbit Unversitas Diponegoro [Diponegoro University Publishing Agency].

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of Educational Measurement*, *44*(4), 325–340. https://doi.org/10.1111/j.1745-3984.2007.00042.x

Grigorovitch, A. (2014). Children's misconceptions and conceptual change in Physics Education: The concept of light. *Journal of Advances in Natural Sciences*, *1*(1), 34–39. https://doi.org/10.24297/jns.v1i1.5037

Gurel, D., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics Science and Technology Education*, *11*(5), 989-1008. https://doi.org/10.12973/eurasia.2015.1369a

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Holme, P. (2015). Modern temporal network theory: A colloquium. *European Physical Journal B, 88*, Article 234. https://doi.org/10.1140/epjb/e2015-60657-4

Huda, C., Sulisworo, D., & Toifur, M. (2017). *Analisis buku ajar termodinamika dengan konsep technological pedagogical and content knowledge (TPACK) untuk Penguatan kompetensi belajar mahasiswa* [Analysis of thermodynamics textbooks with technological pedagogical concepts and content knowledge (TPACK) for strengthening student learning competencies]. *Jurnal Penelitian Pembelajaran Fisika*, *8*(1), 1–7. https://doi.org/10.26877/jp2f.v8i1.1330

Istiyono, E., Mardapi, D., & Suparno, S. (2014). *Pengembangan tes kemampuan berpikir tingkat tinggi fisika (pysthots) peserta didik SMA* [Development of higher order thinking physics (pyshots) tests for high school students]. *Jurnal Penelitian dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Kandil İngeç, Ş. (2009). Analyzing concept maps as an assessment tool in teaching physics and comparison with the achievement tests. *International Journal of Science Education*, *31*(14), 1897–1915. https://doi.org/10.1080/09500690802275820

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lane, S., Raymond, M. R., & Haladyna, T. M. (Eds.). (2016). *Handbook of test development* (2nd ed.). Routledge.

Lissitz, R. W., & Samuelsen, K. (2007). Further clarification regarding validity and education. *Educational Researcher*, *36*(8), 482–484. https://doi.org/10.3102/0013189x07311612

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes* [Techniques for preparing test and non-test instruments]. Mitra Cendekia.

Mardapi, D. (2017). *Pengukuran penilaian dan evaluasi pendidikan* [Measurement of educational assessment and evaluation] (2nd ed.). Parama Publishing.

Mayers, A. M., Khoo, S. T., & Svartberg, M. (2002). The Existential Loneliness Questionnaire: Background, development, and preliminary findings. *Journal of Clinical Psychology*, *58*(9), 1183–1193. https://doi.org/10.1002/jclp.10038

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Pearson.

Prahani, B. K., Amiruddin, M. Z., Suprapto, N., Deta, U. A., & Cheng, T. H. (2022). The trend of physics education research during COVID-19 pandemic. *International Journal of Educational Methodology, 8*(3), 517-533. https://doi.org/10.12973/ijem.8.3.517

Pratama, G. S., & Retnawati, H. (2018). Urgency of higher order thinking skills (HOTS) content analysis in mathematics textbook. *Journal of Physics: Conference Series*, *1097,* Article 12147. https://doi.org/10.1088/1742-6596/1097/1/012147

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

Ratnaningdyah, D. (2018). Mengungkap miskonsepsi fisika dengan metode the three-tier test [Revealing physics misconceptions with the three-tier test method]. In *Prosiding seminar nasional program pascasarjana universitas PGRI Palembang* [Proceedings of the national seminar on the postgraduate program Palembang PGRI university] (pp. 533–540). Palembang University.

Retnawati, H. (2014). *Teori respons butir dan penerapannya* [Item response theory and its application] (1st ed.). Nuha Medika.

Retnawati, H. (2016a). *Analisis kuantitatif instrumen penelitian (panduan peneliti, mahasiswa, dan psikometrian)* [Quantitative analysis of research instruments (researcher, student, and psychometrical guide)]. Parama Publishing.

Retnawati, H. (2016b). *Validitas reliabilitas dan karakteristik butir* [The validity of the reliability and characteristics of the item]. Parama Publishing.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*(1), 1–25. https://doi.org/10.1037/0033-2909.132.1.1

Rosyid, Jatmiko, B., & Supardi, Z. A. I. (2013). *Meningkatkan hasil belajar fisika menggunakan model orientasi IPA (PBL berbasis multirepresentasi) pada konsep mekanika di SMA* [Improving physics learning outcomes using a science orientation model (multi-representation-based PBL) on the concept of mechanics in high school]. *Jurnal Pendidikan Fisika*, *2*(3), 1–12. https://l24.im/6VmLE

Saepuzaman, D., & Karim, S. (2016). Desain pembelajaran student's conceptual construction guider berdasarkan kesulitan mahasiswa calon guru fisika pada konsep gerak parabola [Student's conceptual construction guider learning design based on difficulty of prospective physics teacher students on parabolic motion concept]. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, *2*(2), 79-86. https://doi.org/10.21009/1.02211

Saepuzaman, D., Retnawati, H., & Istiyono, E. (2021). Can innovative learning affect student HOTS achievements?: A meta-analysis study. *Pegem Journal of Education and Instruction*, *11*(4), 290–305. https://doi.org/10.47750/pegegog.11.04.28

Saepuzaman, D., Utari, S., & Nugraha, M. G. (2019). Development of basic physics experiment based on science process skills (SPS) to improve conceptual understanding of the preservice physics teachers on Boyle's law. *Journal of Physics: Conference Series*, *1280*, Article 052076. https://doi.org/10.1088/1742-6596/1280/5/052076

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Silung, S. N. W., Kusairi, S., & Zulaikah, S. (2016). *Diagnosis miskonsepsi siswa sma di kota malang pada konsep suhu dan kalor menggunakan three tier test* [Diagnosis of misconceptions of high school students in malang city on the concept of temperature and heat using the three tier test]. Jurnal Pendidikan Fisika dan Teknologi, *2*(3), 95-105. https://doi.org/10.29303/jpft.v2i3.295

Subali, B., & Suyata, P. (2012). *Pengembangan item tes konvergen dan divergen dan penyelidikan validitasnya secara empiris* [Development of convergent and divergent test items and investigation of their empirical validity]. Diandra Pustaka Indonesia.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan RASCH pada assessment pendidikan* [Application of RASCH modeling in educational assessment]. Trim komunikata Bandung

Suwarto. (2013). *Belajar tuntas, miskonsepsi, dan kesulitan belajar* [Complete learning, misconceptions, and learning difficulties]. *Jurnal Hasil Riset, 22*(1), 85 – 95. https://l24.im/XrUQIp

Tiandho, Y. (2018). *Miskonsepsi gaya gesek pada mahasiswa* [Frictional force misconceptions on undergraduate student]. *Jurnal Pendidikan Fisika dan Keilmuan*, *4*(1), 1-9. https://doi.org/10.25273/jpfk.v4i1.1814

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

Yang, T. C., Chen, S. Y., & Hwang, G. J. (2015). The influences of a two-tier test strategy on student learning: A lag sequential analysis approach. *Computers and Education*, *82*, 366–377. https://doi.org/10.1016/j.compedu.2014.11.021

Yanuike, A. W., Setyarsih, W., & Kholiq, A. (2017). *Penggunaan Phet simulation dalam ecirr untuk mereduksi miskonsepsi siswa pada materi fluida dinamis* [The use of Phet simulation in ecirr to reduce students' misconceptions on dynamic fluid materials]. *Inovasi Pendidikan Fisika*, *5*(3), 161–164. https://l24.im/nl0b

Zukhruf, K. D., Khaldun, I., & Ilyas, S. (2017). Remediasi miskonsepsi dengan menggunakan media pembelajaran interaktif pada materi fluida statis [Remediation of misconceptions by using interactive learning media on static fluid materials]. *Indonesian Journal of Science Education/Jurnal Pendidikan Sains Indonesia, 4*(1), 56–68. https://l24.im/mFY

Zulfikar, A., Saepuzaman, D., Novia, H., Setyadin, A. H., Jubaedah, D. S., Sholihat, F. N., Muhaemin, M. H., Afif, N. F., Fratiwi, N. J., Bhakti, S. S., Amalia, S. A., Hidayat, S. R., Nursani, Z., Hermita, N., Costu, B., & Samsudin, A. (2019). Reducing eleventh-grade students' misconceptions on gravity concept using PDEODE∗E-based conceptual change model. *Journal of Physics: Conference Series*, *1204*, Article 012026. https://doi.org/10.1088/1742-6596/1204/1/012026