Characteristics of Fundamental Physics Higher-order Thinking Skills Test Using Item Response Theory Analysis

Duden Saepuzaman[1], Edi Istiyono[2*], Haryanto[3]

[1]Graduate School, Educational Research and Evaluation program, Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta 55281, Indonesia (*as student doctoral program*)

[1]Education Department, Universitas Pendidikan Indonesia, Jl. Dr. Setiabudhi No. 229, Bandung 40154, Indonesia (*as Lecturer*)

[2,3]Universitas Negeri Yogyakarta, Jl. Colombo No. 1, Yogyakarta 55281, Indonesia

Abstract

HOTS is one part of the skills that need to be developed in the 21st Century . This study aims to determine the characteristics of the Fundamental Physics Higher-order Thinking Skill (FundPhysHOTS) test for prospective physics teachers using Item Response Theory (IRT) analysis. This study uses a quantitative approach. 254 prospective physics teachers at West Java and Banten, Indonesia. Data were collected through tests to respond to the FundPhysHOTS test. The FundPhysHOTS test instrument consists of 26 items and Two-Tier Multiple-Choice (TTMC) in form with a polytomous score of 4 categories (1,2,3 and 4). Data analysis includes two stages: the IRT assumption test and then continued with IRT analysis to determine item and ability parameters. The results show that the data are unidimensional and local independence based on empirical data so that the IRT assumption is fulfilled. The IRT analysis used is the generalized partial credit model (GPCM). The results of the item parameter analysis show that all items have good discriminatory power parameters ($0.394 < a_i < 1.397$) and are classified as good. The difficulty level analysis showed that almost all items had good step parameters ($b_1 < b_2 < b_3$), with the mean of difficulty index (b mean) in the range of -0.332 to 0.144 and categorized as moderate. The ability (Θ) of prospective physics teachers is in the range of $-3.976 < \Theta < 1.646$. Information function analysis shows that the FundPhysHOTS instrument is reliable in measuring ability in this range. This study provides an overview of the analysis of test quality using the polytomous IRT analysis for the benefit of developing the test and developing further studies.

Keywords: FundPhysHOTS, Item Response Theory, Prospective Physics Teachers.

Introduction

Science and technology in this century are developing very rapidly and have influenced all aspects of life, including education (Bond et al., 2019; Hariharasudan & Kot, 2018; Teräs et al., 2020). One of the skills needed today and in the future is HOTS (Hasan & Pardjono, 2019; Pratama & Retnawati, 2018; Suhendro et al., 2021). It is not surprising that HOTS is part of learning goals or achievements in schools and schools or higher education (Jansen et al., 2019; Maphalala & Adigun, 2020; Wang & Zheng, 2021). HOTS learning will prepare students to challenge the current and future workforce (Mitani, 2021; Yeung, 2015). Other researchers state that learning that trains HOTS answers complex questions and solutions to a case or problem through learning (Sukatiman et al., 2020; Tyas & Naibaho, 2021).

In its implementation, learning held in schools refers to the curriculum (Kim & Jung, 2019). Generally, the learning objectives listed in the curriculum are various competencies that must be trained and possessed by students (José Sá & Serpa, 2018; Mukminin et al., 2019). One of the next generation's competencies is higher-order thinking skills (Tondeur et al., 2019). Regarding this demand, for example, in Indonesia's curriculum, it is natural for the government to try to improve the quality of education through various curriculum reforms, starting from the direction of the

recommended learning approach to the assessment of learning outcomes—for example, the 2013 curriculum revision (Warman et al., 2021). Implementation of the 2013 curriculum is an effort to improve the quality of education-oriented towards achieving students' higher-order thinking skills (Iriyanti & Darwis, 2021). Teachers are still not used to making and using HOTS questions in measuring student learning outcomes. One of the teacher's difficulties is to ensure that the questions they make actually measure students' HOTS (Walsh et al., 2007). Thinking skills categorized as higher-order thinking skills include critical thinking skills and creative thinking (Conklin, 2012; King et al., 2010; Krulik & Rudnick,

1999; Presseisen, 1988). These skills are not foreign terms in the learning process; they have even become targets and part of the learning objectives in each subject (Jailani et al., 2018). An assessment is used to determine the HOTS achievement of students. Assessment is a process of gathering information related to learning objectives or achievements (Kumar et al., 2016). One of the most widely used assessment efforts is in the form of a test instrument. The test instruments commonly used are multiple-choice questions or descriptions, each with advantages and disadvantages. Multiple-choice questions are the most widely used because they are easy to apply and analyze. Multiple-choice questions are often criticized for only assessing superficial memorization or simple facts because they do not allow test takers to explain or justify their answers (Nichols & Sugrue, 1999; Songer et al., 2009). Although in some cases, this weakness can be reduced (Hestenes et al., 1992; Xiao et al., 2018).

The development of reasoned multiple-choice questions (reasoning multiple choice) measures high-level abilities or skills (Liu et al., 2011; Xiao et al., 2018). Suppose the inclusion of reasons at the second level of the two-tier choice question form can improve higher-order thinking skills and see the ability of test-takers to give reasons (Cullinane & Liston, 2011). So that in choosing the answers, the test takers must think about the reasons that match the answer choices, directly the thinking process determines the right reasons to train the higher-order thinking skills of the test takers. In addition, it can be seen that the lack of quality assessment is due to the selection of multiple-choice test models commonly used to measure low-level thinking skills (Istiyono et al., 2014). Multiple-choice tests must be modified to measure higher-order thinking skills (Brookhart, 2010). One of the efforts is making a two-tier instrument, often called a two-tier multiple-choice (TTMC) (Istiyono et al., 2020). Regarding TTMC scoring, an alternative approach that can be used is the item response theory approach for polytomous scoring.

Apart from the form of the test instrument, another element that must be considered in the assessment is to seek and ensure that the assessment results accurately describe students' abilities. An assessment is accurate if the assessment results contain the smallest possible error or error. To get results that accurately describe students' ability, the quality of the test instrument must be valid, reliable, and have good item parameters. For this purpose, two approaches can be used to estimate item parameters, namely classical test theory and item response theory. Classical test theory is seen as having weaknesses. The most notable weakness of classical test theory is that examinee characteristics and test characteristics cannot be separated, each of which can only be interpreted in another context (Hambleton et al., 1991). That is, the test only determines the ability of the examinee. When the test is difficult, the examinee will have the low ability. Otherwise, the

examinee will appear to have a higher ability when the test is easy. In other words, item parameters are highly dependent on the subject/test taker and vice versa. The characteristics of the items will change when the examinees change, and the characteristics of the examinees will change when the items change. In this case, classical test theory cannot be used as a standard because the assessment results depend on the test taker's subjects.

Item response theory is a solution to overcome the weaknesses in classical test theory because item response theory has the concept of releasing the link between items and samples or test takers. The characteristics/ability of the examinees will remain the same even if they work on items with different characteristics. Conversely, the characteristics of the items will remain the same even if examinees perform them with different abilities. In addition, the item response theory is based on items/items no longer on test kits. Item response theory rests on two postulates: (a) test takers' performance on test items can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) when the ability increases, the respondent's probability of answering correctly for an item increases. The function of item response theory can be applied when the model used has a good fit with the test (Hambleton et al., 1991). Item parameter estimation could be disrupted when the model does not match the data (Stone & Zhang, 2003). In the IRT approach with polytomous scoring, several models are known, including the Graded Response Model (GRM), Partial Credit Model (PCM), and Generalized Partial Credit Model (GPCM). This study uses GPCM analysis. GPCM model is suitable for analyzing multiple-choice data (Si & Schumacker, 2004). The same thing is also reinforced by the opinion of Retnawati (Retnawati, 2011), which states that the GPCM is the most suitable model for analyzing test results with the polytomous scoring model because this item is the score in a tiered category. Still, the difficulty index in each step is not ordered; a step can be more difficult than the next step.

Based on the background, this study focused on determining the characteristics of the FundPhysHOTS test for prospective physics teachers using polytomous Item Response Theory (IRT), namely GPCM analysis. The analysis includes testing the IRT assumption and determining item parameters and abilities.

METHOD

Research Design

This study is descriptive research with a quantitative approach. This study describes the quality of the items that make up the instrument based on quantitative data obtained from data analysis derived from prospective physics teacher responses to the FundPhysHOTS test.

Population and Sample/ Study Group/Participants

The subjects of this study were 254 prospective physics teacher students in two universities, which produce prospective teachers graduate. These 254 students were 138 prospective physics teachers at one of the universities in West Java province and 116 prospective physics teachers at one of the universities in Banten province.

Data Collection Tools

The FundPhysHOTS was developed based on the HOTS indicator, which refers to critical and creative thinking skills

(Saepuzaman, Retnawati & Istiyono, 2021) of 7 aspects. The aspects and indicators are is presented in Table 1.

FundPhysHOTS contains HOTS testing on fundamental physics concepts, namely on the material of one-dimensional motion kinematics, two-dimensional motion kinematics, and particle dynamics. The test instrument is in Two-Tier Multiple Choice (TTMC) with four assessment categories, as shown in Figure 1. The scoring criteria are presented in Table 2 (Istiyono et al., 2020).

The validity and reliability of this instrument have been proven in previous studies. Content validity using V Aiken,

Tabel 1: *Aspect and Indicator of FundPhysHOTS instrument test*

Aspect	Indicator
Elementary Clarification	• Identify/formulate questions based on events in everyday life
	• Analyze statements and determine the similarities or differences of a given event
Basic Support	• Examine/examine the parts that can be considered trustworthy (or untrustworthy) based on argumentative texts, advertisements, or experiments and their interpretations, and give reasons
	• Expressing reasons based on observations of an event
Inference	• Interpret statements and can clarify data
	• Generalize (find patterns) based on the trend of existing data
Strategy And Tactics	• Solving problems using definitions
	• Formulate alternatives to solutions
Fluency	• Answer with some answers or facts
	• Seeing the faults and weaknesses of an object or situation
Flexibility	• Provide interpretation of an image, story, or issue
	• Thinking of ways or points of view to solve problems
	• Classify things according to different divisions (categories)
Elaboration	• Develop or enrich the ideas of others
	• Trying out/testing new things by trial

Figure 1: Example of Two-Tier Multiple Choice (TTMC) questions

Table 2: Scoring Criteria

Score	Criteria
4	Answers to questions and reasons are correct
3	The answer to the question is wrong, but the reason is correct
2	The answer to the question is correct, but the reason is wrong
1	The answer to the question and the reason are wrong

Table 3: KMO dan Bartlett's Test

KMO dan Bartlett's		Value
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.917
Bartlett's Test of Sphericity	Approx. Chi-Square	3098.119
	df	325
	Sig.	0.000

the value of V = 0.867 (valid). Construct validity and reliability analysis using Confirmatory Factor Analysis (CFA). The results indicated that most of the items developed had good construct validity (with a loading factor (λ) value of more than 0.4). The construct reliability of this test instrument belongs to the reliable category (reliability construct coefficient value 0.514, more than 0.5; and Cronbach Alpha value is 0.86, more than 0.80) (Appendix 1). These results prove that the FundPhysHOTS test instrument is valid and reliable to measure the HOTS of prospective physics teachers.

Data Analysis

Data analysis was carried out in two stages: testing the IRT assumptions using SPSS software, determining the item parameters, and using the R program.

FINDINGS

Assumption Test

The first analysis in the IRT approach is the dimensionality test. The purpose of investigating whether the test instrument is unidimensional or multidimensional). Unidimensional means that each item only measures one ability. On the other hand, multidimensional implies that some or all items measure more than one dimension (Retnawati, 2014). The dimensional test in this study was proven through factor analysis using SPSS. Factor analysis was conducted by first conducting a feasibility test analysis, namely the KMO-MSA test and the Barlett test. The KMO-MSA test aims to see the adequacy of the sample, while the Barlett Test serves to prove the homogeneity of the data. Factor analysis can be continued if the Kaiser Meyer Olkin (KMO)-MSA value> 0.5 and the Barlett significant test <0.05 (Hair et al., 2009; Widarjono, 2020). Based on the student response data, the KMO-SMA and Barlett scores are obtained as presented in Table 3. Table 3 shows that the sample used has met the sample adequacy requirements (KMO-MSA = 0.917 > 0.5), and the data is homogeneous (Barlett test <0.05), so that factor analysis can be performed.

A test is considered unidimensional if it measures only one dominant dimension (Widarjono, 2020). The number of factors created can be determined by the presence of an eigenvalue greater than one, which is the indicator factor (Retnawati, 2014; Widarjono, 2020). Factor analysis identified five components with eigenvalues greater than one (Appendix 2).
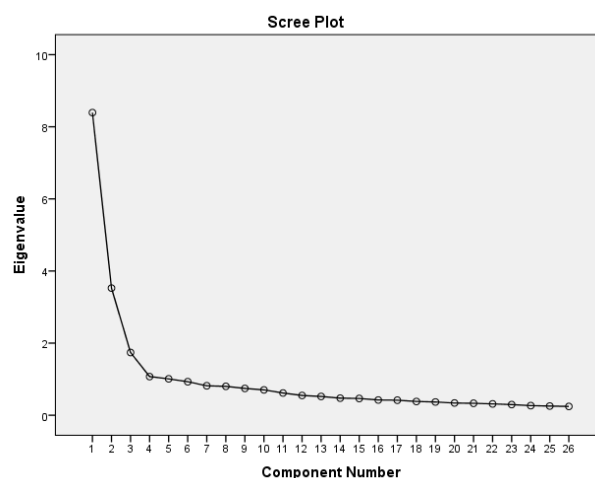
Figure 2: Scree plot

It shows that the 26 FundPhyHOTS test items consist of 5 factors. The analysis results show that factor 1 is the dominant factor because its eigenvalue is 8,393, larger than the others or the most dominant, implying that FundPhysHOTS is unidimensional. Statistical analysis also shows an eigenvalue of 8.393, where the result is more than two times the eigenvalue of the second factor with a percentage variance of 32.82%. Cumulatively, the percentage of the five factors is 60.542, indicating that the five existing components explain 65.546%. The cumulative percentage of 60.542% has met the minimum conditions for the cumulative value of taking the right number of variables, which is 50% (Widarjono, 2020; Retnawati et al., 2017; Wells & Purwono, 2009). It further strengthens that the FundPhysHOTS test instrument is unidimensional.

The dimensions recorded in the data can be proven on the scree plot, especially the steep numbers. The number of steps indicates the number of dimensions/factors, while the slope of the change in the eigenvalues does not indicate any dimensions (Widarjono, 2020). Therefore, unidimensionality can also be shown from the next scree plot. The test is considered unidimensional when components 1 and 2 in the scree plot have a sufficiently high distance (Furr & Bacharach, 2008). According to the scree plot in Figure 2, component 1 is located far from component 2, while component 2 is located quite close to component 3 and other components. Moreover, as illustrated in Figure 2, the eigenvalues begin skewed with

the third component. It further confirms that the instrument FundPhysHOTS is unidimensional.

Another assumption test is local independence. This local independence assumption will be fulfilled if the participant's answer to one item does not affect the participant's response to the other items (Retnawati, 2014). According to De Mars (Ockey, 2013), local independence can also be detected by proving the unidimensional assumption. It means that if the unidimensional assumption is met, the local independence assumption is also met. In this study, the unidimensional assumption has been fulfilled so that the local independence test has also been fulfilled.

Fit Model

The next stage in the IRT analysis is the suitability of the analytical model, namely GPCM, with empirical data obtained

Table 4: Recapitulation of item fit with GPCM model

Item	statistics	df	RMSEA	p-Value	Remarks
Q_1	47.02	51.00	0.00	0.63	Fit
Q_2	54.47	50.00	0.02	0.31	Fit
Q_3	38.83	46.00	0.00	0.76	Fit
Q_4	52.43	45.00	0.03	0.21	Fit
Q_5	54.74	46.00	0.03	0.18	Fit
Q_6	58.06	50.00	0.03	0.20	Fit
Q_7	61.77	43.00	0.04	0.03	Fit
Q_8	41.88	42.00	0.00	0.48	Fit
Q_9	46.07	43.00	0.02	0.35	Fit
Q_10	33.62	43.00	0.00	0.85	Fit
Q_11	43.05	45.00	0.00	0.55	Fit
Q_12	59.02	52.00	0.02	0.23	Fit
Q_13	50.11	49.00	0.01	0.43	Fit
Q_14	59.58	50.00	0.03	0.17	Fit
Q_15	36.07	45.00	0.00	0.83	Fit
Q_16	44.06	47.00	0.00	0.60	Fit
Q_17	47.29	47.00	0.00	0.46	Fit
Q_18	52.44	52.00	0.01	0.46	Fit
Q_19	46.03	49.00	0.00	0.59	Fit
Q_20	37.81	46.00	0.00	0.80	Fit
Q_21	59.24	46.00	0.03	0.09	Fit
Q_22	71.09	51.00	0.04	0.03	Fit
Q_23	41.08	47.00	0.00	0.72	Fit
Q_24	59.26	47.00	0.03	0.11	Fit
Q_25	66.55	49.00	0.04	0.05	Fit
Q_26	64.91	45.00	0.04	0.03	Fit

by the researcher. The fit test model will use the p.S_X2 method. This method is suitable for instruments with few items (Arlinwibowo, Retnawati & Kartowagiran, 2021). A data or empirical evidence is said to fit the model if the p-value of p.S_X2 is greater than 0.05 or RMSEA, S_X2 approaches zero (Chalmer & Ng, 2017). The suitability of each item with the model presented in Table 4. Based on the p-value and RMSEA shows that the FundPhysHOTS (for all items), empirically, has a match with GPCM analysis.

Item Parameters

Since all the items in FundPhysHOTS fit the GPCM, the next step is to estimate the item parameters, including discriminatory power (a) and item difficulty index (b). The results of the GPCM analysis resulted in the item parameters presented in Table 5.

Table 5 shows that the overall item discriminatory power is within the parameters ranges 0.394 to 1.397 (good), and the difficulty index of all items had good step parameters (b1 < b2 < b3), with b, mean in the range of -0.332 to 0.144. So that the whole item can be accepted as a good item, because the value range of Discriminatory power (a) between 0 and 2 (0 < a < 2) and the difficulty index (b) between -2 and +2 (-2 <b <+2) (Maryani et al. al , 2022 ; Widarjono , 2015 ; Retnawati 2014).

The relationship between the probability of answering correctly for each ability or step parameter is presented in the Item Characteristic Curve (ICC) (du Toit, 2003; Retnawati, 2014). For example, ICC in four for item number 3 (Q _3) is presented in Figure 3.

Figure 2 shows the step parameters difficulty index which is indicated by the intersection of the curves. For example, the intersection point between the blue line and the purple line (1 and 2) states the minimum opportunity and ability, often referred to as the first step parameter (b1). Step parameter (b1) represents the step parameter from score 1 to score 2, -1.198. The same interpretation for the next step parameters, b2 and b3, are -0.323 and 0.821. ICC for all items is presented in Figure 4.

The next analysis related to the instrument profile is to look at the value of the test information (I). The analysis will be used to determine the instrument's suitability with students based on their abilities. Figure 5 will display a graphic image of the test information value.

Ability

Estimation of prospective physics teachers' ability to answer the FundPhysHOTS test is done after the model fit analysis, and the estimation of the item parameters are determined. The ability determination method used in this study is the maximum likelihood estimation (MLE). Statistical descriptions related to capability parameters are presented in Table 6.

Table 5: Item Parameters Analysis of FundPhysHOTS instrument test

| Item	Discrimantory Power		Difficulty Index					Conclusion
	a	Remarks	b1	b2	b3	b mean	Remarks	
Q_1	0.507	Good	-1.734	-0.37	0.478	-0.280	Good	Accepted
Q_2	0.510	Good	-1.536	-0.553	1.061	-0.130	Good	Accepted
Q_3	1.136	Good	-1.198	-0.323	0.821	0.109	Good	Accepted
Q_4	0.899	Good	-1.938	-0.159	1.163	-0.009	Good	Accepted
Q_5	1.171	Good	-1.027	-0.501	0.886	0.132	Good	Accepted
Q_6	0.655	Good	-2.04	-0.114	0.536	-0.241	Good	Accepted
Q_7	1.135	Good	-1.912	-0.187	0.898	-0.017	Good	Accepted
Q_8	1.397	Good	-1.431	-0.245	0.856	0.144	Good	Accepted
Q_9	1.295	Good	-1.31	-0.394	0.881	0.118	Good	Accepted
Q_10	1.283	Good	-1.557	-0.242	0.825	0.077	Good	Accepted
Q_11	1.115	Good	-1.418	-0.210	0.659	0.037	Good	Accepted
Q_12	0.394	Good	-2.374	-1.129	1.843	-0.317	Good	Accepted
Q_13	0.601	Good	-1.544	-0.958	0.688	-0.303	Good	Accepted
Q_14	0.616	Good	-1.824	-0.415	0.297	-0.332	Good	Accepted
Q_15	0.884	Good	-1.481	-0.38	0.869	-0.027	Good	Accepted
Q_16	0.595	Good	-1.954	-0.626	1.405	-0.145	Good	Accepted
Q_17	0.616	Good	-1.582	-0.717	1.35	-0.083	Good	Accepted
Q_18	0.439	Good	-0.834	-1.483	0.937	-0.235	Good	Accepted
Q_19	0.596	Good	-1.521	-0.555	1.042	-0.110	Good	Accepted
Q_20	0.836	Good	-1.748	-0.38	0.891	-0.100	Good	Accepted
Q_21	0.887	Good	-1.392	-0.878	0.865	-0.130	Good	Accepted
Q_22	0.434	Good	-2.081	-0.374	1.179	-0.211	Good	Accepted
Q_23	0.691	Good	-1.284	-0.876	1.053	-0.104	Good	Accepted
Q_24	0.748	Good	-1.521	-0.083	0.917	0.015	Good	Accepted
Q_25	0.781	Good	-1.676	-0.661	0.587	-0.242	Good	Accepted
Q_26	0.872	Good	-1.882	-0.37	1.249	-0.033	Good	Accepted
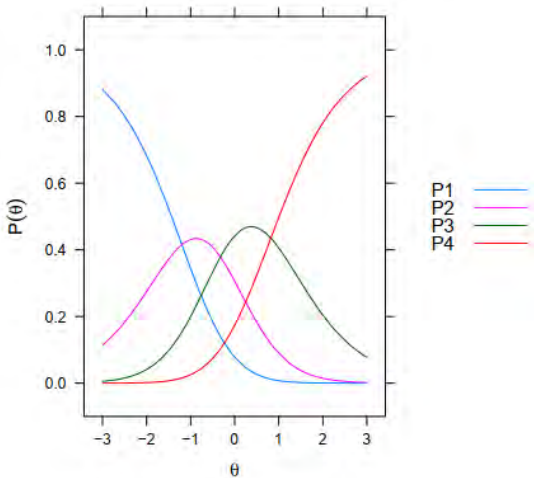
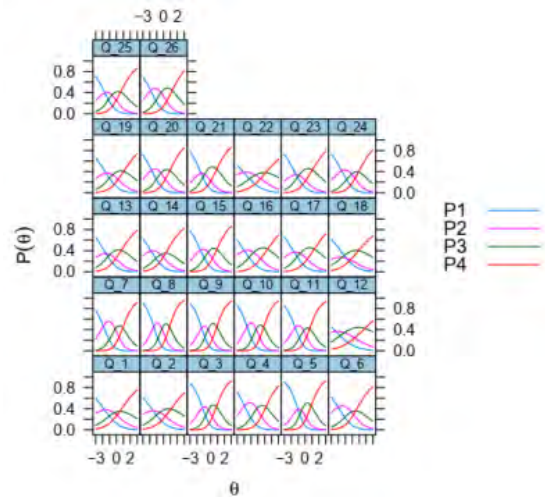**Fig. 3: Item Characteristic Curve (ICC) item 5 (Q _3)**

Figure 4. ICC for all item

The complete distribution of the abilities of prospective physics teachers is presented in the appendix. If presented in a histogram, the general distribution is presented in Figure 6.

The ability and difficulty level distribution map of the same scale can be seen from the wright map (Chan, Looi, & Sumintono, 2021). The Wright Map provides a picture of an exam by placing the difficulty of the exam items on the same measurement scale as the ability of the individuals. The Wright Map is organized as two vertical histograms. The left side shows the person and the right side shows items. The left side of the map shows the distribution of the measured ability of the candidates from most able at the top to least able at the

bottom. The items on the right side of the map are distributed from the most difficult at the top to the least difficult at the

Table 6: Statistic Description ability ( ) estimation

Stats	Value
Number of participant	254
Mean	-0.02
Standard Deviation of Thetas ( )	1.087
Marginal Reliability	0.916
Mean of Standard Errors	0.306
Maximum	1.646
Minimum	-3.976

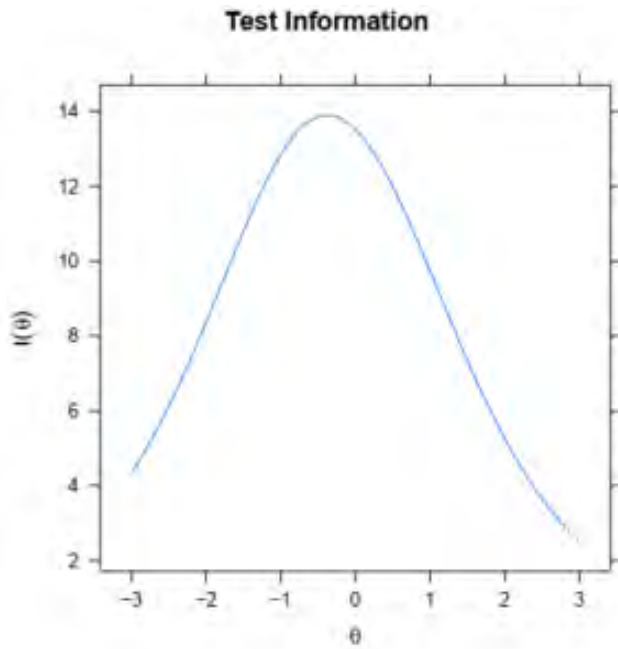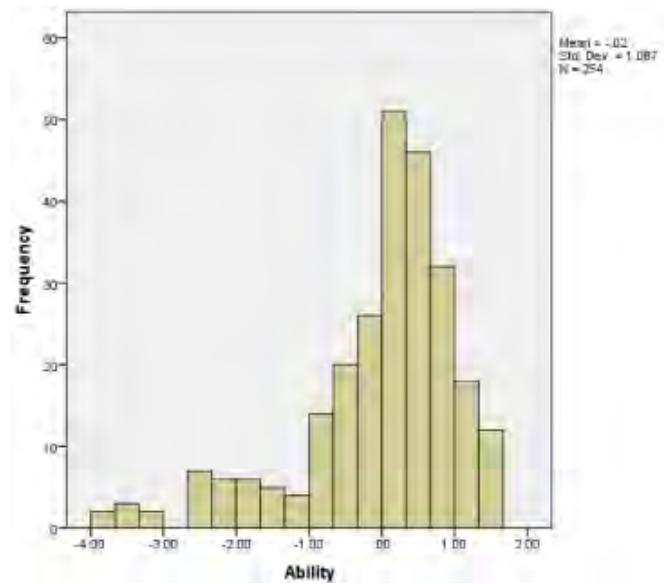**Fig. 5: Test Information Function Curve**

Fig. 6: The histogram of prospective physics teacher abiity
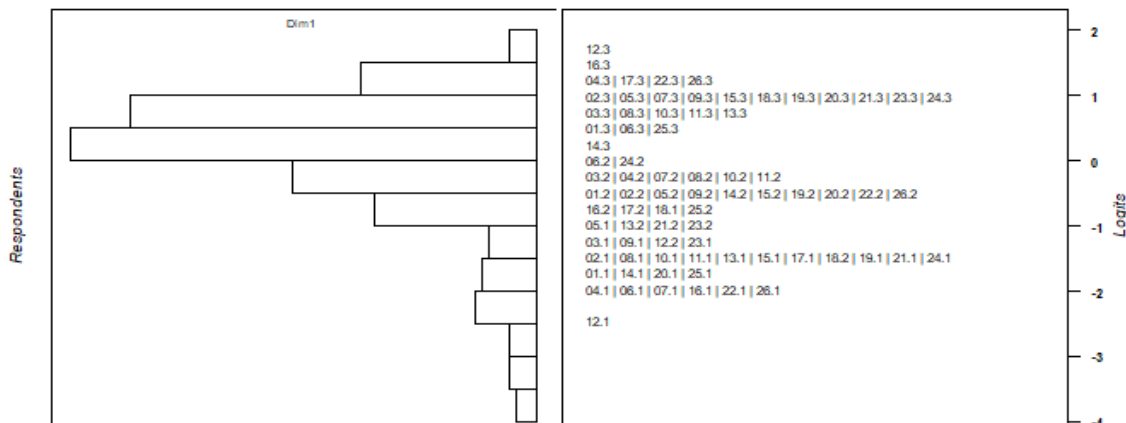

Figure 7: Wright Map Analysis

bottom. An overview of the wright map is presented in Figure 7. Based on figure 7, the item with the highest difficulty level is 12.3 (item no 12 in category 3), while the easiest is 12.1 (item no 12 in category 1). The picture on the right shows the frequency of students' abilities with the same logit scale as the item's difficulty level—the distribution of students' abilities with a logit scale of 0 to 1.

Discussion

The analysis begins with the item response theory assumptions, are unidimensional assumptions, and local independence. Unidimensional assumptions are proven by exploratory factor analysis using the SPSS program. Based on the factor analysis results carried out using the Bartlett test, the KMO value was 0.881. The KMO value is greater than 0.5, meaning that the 254 samples used in this study were sufficient to analyze. Next, by reviewing the eigenvalues, it can be seen that the first component is the dominant factor with an eigenvalue of 8.393, which can explain 32.82 % of total variances. Thus, the FundPphysHOTS instrument fulfills unidimensionality. Wells and Purwono (Apino & Retnawati, 2016) confirm the amount of the explained variance presentation; if the value is greater than 20%, then the device being measured contains a single dimension or is unidimensional. The unidimensionality is fulfilled, meaning FundPhysHOTS only measures one dimension (Hambleton et al., 1991). The dimension or ability measured here is HOTS. The unidimensional proof is also reinforced by scree plot data, which shows that one component has a dominant steepness compared to other components (Furr % Bacharach, 2008).

The fulfillment of the unidimensional assumption based on the results above shows that the assumption of local independence is also fulfilled (Retnawati, 2014). It is because the data held is unidimensional. The response given by the test taker to an item is independent or does not affect the test taker's answer to other items. Test-takers ability is independent of the items of the FundPhysHOTS instrument test.

Based on the data in table 3, it appears that the fittest or provide information on each item for FundPhysHOTS instruments is GPCM. These results are in line with the opinion of Si (Si & Schumacker, 2004), which states that the GPCM model is suitable for analyzing multiple-choice data. The same thing is also reinforced by the opinion of Retnawati (2011: 2), which states that the GPCM is the most suitable model for analyzing test results with the polytomous scoring model because this item is get score in a tiered category, but the difficulty index in each step is not ordered, a step can be more difficult than the next step. Istiyono (2020) asserts that using GPCM to analyze multiple-choice tests is a fair alternative assessment model in learning (Istiyono et al., 2020).

Overall item discriminatory power is within the parameters range 0.394 to 1.397 (good) , and the difficulty index of all items had good step parameters (b1 < b2 < b3), with b mean in the range of -0.332 to 0.144. So that the whole item can be accepted as a good item, because the value range of Discriminatory power (a) between 0 and 2 (0 < a < 2) and the difficulty index (b) between -2 and +2 (-2 <b <+2) (Maryani et al. al , 2022 ; Widarjono , 2015 ; Retnawati 2014). These characteristics become very important, considering the role of the test instrument must measure test taker's ability as accurately as possible, distinguishing test-takers whose abilities are low, medium, or high.

Further analysis related to the quality of grain parameters can be seen from the step parameters. All items produce 3 step parameters symbolized by bi (Ostini & Nering, 2006). The value of bi is the intersection of the mn and mn+1 category curves (Embretson & Reise, 2000). bi refers to a certain minimum ability to enter a higher point category (Retnawati, 2014). The data in Table 4 shows that the values of b1, b2, and b3 for each item have a good order, namely b1 < b2 < b3 (Van der Linden, 2017). It is reinforced by the ICC curve presented in Figure 4. Thus, the difficulty level of each item is of good quality and can represent the HOTS ability of prospective physics teachers well.

Theta, denoted by ( ), is a psychometric term in item response theory that indicates the test taker's ability to be measured, in this case, is the ability of HOTS. Based on the value of the information function shown on the di curve in Figure 5, the FundPhysHOTS test instrument accurately measures students' ability (θ) between intervals of -3 to 3. It means that FundPhysHOTS can provide accurate information regarding students' HOTS abilities for various abilities ranging from low, medium , and high. The most accurate information is obtained for those with moderate abilities (in the around zero). It can be seen from the peak of the information function.

HOTS abilities of prospective physics teachers are quite diverse. It can be seen from the statistical description in Table 5, which shows that theta values vary from -3.976 to 1.646. It is clearer from the histogram in Figure 6. Although it varies, the distribution of this ability tends to be dominated by moderate ability (around zero). The mean value reinforces this for theta, which is -0.02. Analysis related to the description of the HOTS ability distribution of prospective physics teachers and the distribution of the difficulty level of each step/step parameter can be seen from the wright map in Figure 7. It can be seen that the most difficult item is item number 12 (Q-12) in step 3 (b3=) 1843) or the top right wright map it says "12.3". While the easiest question is number 12 (Q-12) in step 1 (b1) .12.1, with a value of b1 on a logit scale of -2.374. This condition corresponds to the data in table 4. Another analysis can be seen that the dominance of the teacher candidate's ability is on the logit scale of zero (-0.02) or moderate ability. This ability corresponds to the same logit scale as item number 14 (Q_14) in parameter step 3 (b3), or in the wrigth map it says "14.3".

CONCLUSION

The result of the analysis characteristics of FundPhysHOTS using IRT show that the data are unidimensional and local independence so that the IRT assumption is fulfilled. The IRT analysis using GPCM show that all items of FundPhysHOTS have good discriminatory power parameters ($0.394 < a_i < 1.397$) and are classified as good. The difficulty level analysis showed that almost all items had good step parameters ($b_1 < b_2 < b_3$), with b_{mean} in the range of -0.332 to 0.144 and was categorized as moderate. The ability ( ) of prospective physics teachers is in the range of $-3.976 <$ < 1.646. Information function analysis shows that the FundPhysHOTS instrument is reliable in measuring ability in this range.

SUGGESTION

As a continuation of this research, studies on measuring HOTS in more complex aspects can be investigated, not only focusing on developing HOTS from a critical and creative perspective like this study. This study provides an overview of the analysis of test quality using IRT polytomous GPCM analysis for the benefit of developing tests and developing further studies. So that in practice, this analysis can also be carried out in other fields besides what has been done in this research.

LIMITATION

This study has limitations in terms of the developed HOTS indicators. In this study, the HOTS indicators only cover aspects and indicators on critical and creative thinking skills as part of the HOTS.

ACKNOWLEDGMENTS

Conflict of Interest: The author has no conflict of interest through this study

REFERENCES

Apino, E., & Retnawati, H. (2016). Creative Problem Solving to Improve Students' Higher Order Thinking Skills in Mathematics Instructions. Proceeding of 3Rd International Conference on Research, Implementation and Education of Mathematics and Science, May, 339–346.

Arlinwibowo, J., Retnawati, H., & Kartowagiran, B. (2021). Item Response Theory Utilization for Developing the Student Collaboration Ability Assessment Scale in STEM Classes. Journal homepage: http://iieta. org/journals/isi, 26(4), 409-415.

Bond, M., Zawacki-Richter, O., & Nichols, M. (2019). Revisiting five decades of educational technology research: A content and authorship analysis of the British Journal of Educational Technology. British Journal of Educational Technology, 50(1), 12–63. https://doi.org/10.1111/bjet.12730

Brookhart, S. M. (2010). How to Assess Higher-Order Thinking Skills in Your Classroom advances. In Journal of Education (Vol. 1, Issue 18). ASCD. www.ascd.org/memberbooks

Chan, S. W., Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A rasch model measurement analysis. Journal of Computers in Education, 8(2), 213-236.

Conklin, W. (2012). Higher-Order Thinking Skills to Develop 21st Century Learners. In Shell Education. Shell Educational Publishing, Inc.

Cullinane, A., & Liston, M. (2011). Two-tier Multiple Choice Question: An Alternative Method of Formatif Assessment for First Year Undergraduate Biology Students. National Center for Excellence In Mathematics and Education Science Teaching and Learning (NCE-MSTL.

du Toit, M. (2003). IRT from SSI: Bilog-MG, Multilog, Parscale, Testfact. Scientific Software International.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Maheah.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). Analise multivariada de dados. In Bookman. Bookman Editora.

Hambleton, R. K., Shavelson, R. J., Webb, N. M., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. 2.

Hariharasudan, A., & Kot, S. (2018). A scoping review on Digital English and Education 4.0 for Industry 4.0. Social Sciences, 7(11), 227. https://doi.org/10.3390/socsci7110227

Hasan, A., & Pardjono, P. (2019). The Correlation of Higher Order Thinking Skills and Work Readiness of Vocational High School Students. Jurnal Pendidikan Teknologi Dan Kejuruan, 25(1), 52–61. https://doi.org/10.21831/jptk.v25i1.19118

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The Physics Teacher, 30(3), 141–158. https://doi. org/10.1119/1.2343497

Iriyanti, I., & Darwis, M. (2021). Increasing Speaking Skills Through the Drama Method in Class IV Students of SD Unggulan Putra Kaili Permata Bangsa. The 2nd International Conference of Linguistics and Culture (ICLC-2, 128–132.

Istiyono, E, Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (pysthots) peserta didik SMA. Jurnal Penelitian Dan Evaluasi Pendidikan, 18(1), 1–12.

Istiyono, E., Dwandaru, W. S. B., Asysyifa, D. S., & Viana, R. V. (2020). Development of computer-based test in critical thinking skill assessment of physics. Journal of Physics: Conference Series, 1440(1), 12062. https://doi.org/10.1088/1742-6596/1440/1/012062

Istiyono, Edi, Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use. European Journal of Educational Research, 9(1), 91–101. https://doi.org/10.12973/eu-jer.9.1.91

Jailani, Sugiman, Retnawati, H., Bukhori, Apino, E., Djidu, H., & Arifin, Z. (2018). Desain Pembelajaran Matematika Untuk Melatihkan Higher Order Thinking Skills (p. 26).

Jansen, R. S., van Leeuwen, A., Janssen, J., Jak, S., & Kester, L. (2019). Self-regulated learning partially mediates the effect of self-regulated learning interventions on achievement in higher education: A meta-analysis. Educational Research Review, 28, 100292. https://doi.org/10.1016/j.edurev.2019.100292

José Sá, M., & Serpa, S. (2018). Transversal competences: Their importance and learning processes by higher education students. Education Sciences, 8(3), 126. https://doi.org/10.3390/educsci8030126

Kim, Y. C., & Jung, J. H. (2019). Conceptualizing shadow curriculum: definition, features and the changing landscapes of learning cultures. Journal of Curriculum Studies, 51(2), 141–161. https://doi.org/10.1080/00220272.2019.1568583

King, F. J., Goodson, L., & Rohani, F. (2010). Higher order thinking skills: Definition, Teaching Strategies, Assessment. http://goo.gl/su233T.

Krulik, S., & Rudnick, J. A. (1999). Innovative tasks to improve critical and creative thinking skills. In D. L. V Stiff & F. R. Curcio (Eds.), from Developing Mathematical reasoning in Grades K-12 (pp. 138–145). NCTM.

Kumar, H., Kumar, S., Dalabh, M., & Ahmad, J. (2016). Measurement and Evaluation in Education. Retrieved October, 10, 1–248. http://www.ipesp.ac.th/learning/websatiti/chapter9/unit9_1_4.html

Liu, O. L., Lee, H. S., & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. Educational Assessment, 16(3), 164–184. https://doi.org/10.1080/10627197.2011.611702

Maphalala, M. C., & Adigun, O. T. (2020). Academics' experience of implementing e-learning in a south african higher education institution. International Journal of Higher Education, 10(1), 1–13. https://doi.org/10.5430/ijhe.v10n1p1

Maryani, I., Prasetyo, Z. K., Wilujeng, I., & Purwanti, S. (2022). Higher-order thinking test of science for college students using multidimensional item response theory analysis. Pegem Journal of Education and Instruction, 12(1), 292-300.

Mitani, H. (2021). Test Score Gaps in Higher Order Thinking Skills: Exploring Instructional Practices to Improve the Skills and Narrow the Gaps. AERA Open, 7, 23328584211016470. https://doi.org/10.1177/23328584211016470

Montoya, A. K., & Edwards, M. C. (2021). The Poor Fit of Model Fit for Selecting Number of Factors in Exploratory Factor Analysis for Scale Evaluation. Educational and Psychological Measurement, 81(3), 413–440. https://doi.org/10.1177/0013164420942899

Mukminin, A., Habibi, A., Prasojo, L. D., Idi, A., & Hamidah, A. (2019). Curriculum reform in indonesia: Moving from an exclusive to inclusive curriculum. Center for Educational Policy Studies Journal, 9(2), 53–72. https://doi.org/10.26529/cepsj.543

Nichols, P., & Sugrue, B. (1999). The Lack of Fidelity Between Cognitively Complex Constructs and Conventional Test Development Practice. Educational Measurement: Issues and Practice, 18(2), 18–29. https://doi.org/10.1111/j.1745-3992.1999.tb00011.x

Ockey, G. J. (2013). Item response theory. In The Routledge Handbook of Language Testing. Oxford University Press. https://doi.org/10.4324/9780203181287-36

Ostini, R., & Nering, M. L. (2006). Polytomous item response theory models (No. 144). Sage.

Pratama, G. S., & Retnawati, H. (2018). Urgency of Higher Order Thinking Skills (HOTS) Content Analysis in Mathematics Textbook. Journal of Physics: Conference Series, 1097(1), 12147. https://doi.org/10.1088/1742-6596/1097/1/012147

Presseisen, B. Z. (1988). Thinking skills: Meanings and models. In D. A. L. Costa (Ed.), Developing minds: A resource book for teaching thinking (pp. 43–48). ASCD.

Retnawati, H. (2011). Mengestimasi kemampuan peserta tes uraian matematika Dengan pendekatan teori respons butir Dengan penskoran politomus Dengan Generalized partial credit model [Estimating the ability of the participants in the mathematical description test Using the item response theory approach With polytomic scoring With the Generalized partial credit model]. In Prosiding Semnas Penelitian Pendidikan dan Penerapan MIPA. UNY (pp. 53–62).

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

Retnawati, H., Munadi, S., Arlinwibowo, J., Wulandari, N. F., & Sulistyaningsih, E. (2017). Teachers' difficulties in implementing thematic teaching and learning in elementary schools. New Educational Review, 48(2), 201–212. https://doi.org/10.15804/tner.2017.48.2.16

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.

Si, C.-F., & Schumacker, R. E. (2004). Ability Estimation Under Different Item Parameterization and Scoring Models. International Journal of Testing, 4(2), 137–181. https://doi.org/10.1207/s15327574ijt0402_3

Si, C.-F., & Schumacker, R. E. (2004). Ability Estimation Under Different Item Parameterization and Scoring Models. International Journal of Testing, 4(2), 137–181. https://doi.org/10.1207/s15327574ijt0402_3

Songer, N. B., Kelcey, B., & Gotwals, A. W. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. Journal of Research in Science Teaching, 46(6), 610–631. https://doi.org/10.1002/tea.20313

Stone, C. A., & Zhang, B. (2003). Assessing Goodness of Fit of Item Response Theory Models: A Comparison of Traditional and Alternative Procedures. Journal of Educational Measurement, 40(4), 331–352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x

Suhendro, Sugandi, D., & Ruhimat, M. (2021). The Urgency of HOTS-Oriented Learning and Assessment Towards Quality of Education in Facing Indonesia Sustainable Development Goals (SDGs) 2030. Proceedings of the 5th Asian Education Symposium 2020 (AES 2020), 566, 237–250. https://doi.org/10.2991/assehr.k.210715.052

Sukatiman, S., Akhyar, M., Siswandari, & Roemintoyo. (2020). Enhancing higher-order thinking skills in vocational education through scaffolding-problem based learning. Open Engineering, 10(1), 612–619. https://doi.org/10.1515/eng-2020-0070

Teräs, M., Suoranta, J., Teräs, H., & Curcher, M. (2020). Post-Covid-19 Education and Education Technology 'Solutionism': a Seller's Market. Postdigital Science and Education, 2(3), 863–878. https://doi.org/10.1007/s42438-020-00164-x

Tondeur, J., Scherer, R., Baran, E., Siddiq, F., Valtonen, T., & Sointu, E. (2019). Teacher educators as gatekeepers: Preparing the next generation of teachers for technology integration in education. British Journal of Educational Technology, 50(3), 1189–1209. https://doi.org/10.1111/bjet.12748

Tyas, E. H., & Naibaho, L. (2021). Hots Learning Model Improves the Quality of Education. International Journal of Research -GRANTHAALAYAH, 9(1), 176–182. https://doi.org/10.29121/granthaalayah.v9.i1.2021.3100

Van der Linden, W. J. (Ed.). (2017). Handbook of Item Response Theory: Volume 2: Statistical Tools. CRC Press.

Van der Linden, W. J. (Ed.). (2017). Handbook of Item Response Theory: Volume 2: Statistical Tools. CRC Press.

Walsh, G., Murphy, P., & Dunbar, C. (2007). Thinking skills in the early years: A guide for practitioners. Stranmillis University College.

Wang, M., & Zheng, X. (2021). Using Game-Based Learning to Support Learning Science: A Study with Middle School Students. Asia-Pacific Education Researcher, 30(2), 167–176. https://doi.org/10.1007/s40299-020-00523-z

Warman, W., Lorensius, L., & Rohana, R. (2021). Curriculum of Management in Improving the Quality of Catholic School Education in Samarinda City, East Kalimantan, Indonesia. In Budapest International Research and Critics Institute (BIRCI-Journal): Humanities and Social Sciences (Vol. 4, Issue 3). East Kalimantan.

Wells, C. S., & Purwono, U. (2009). Assesing the fit of IRT models to item response data. Makalah Pelatihan Psikometri Kerjasama Pascasarjana UNY dengan USAID.

Widarjono, A. (2020). Analisis multivariat terapan dengan program SPSS, AMOS, dan SMARTPLS.[Applied multivariate analysis with SPSS, AMOS, and SMART PLS programs]

Xiao, Y., Han, J., Koenig, K., Xiong, J., & Bao, L. (2018). Multilevel Rasch modeling of two-Tier multiple-choice test: A case study using Lawson's classroom test of scientific reasoning. Physical Review Physics Education Research, 14(2), 20104. https://doi.org/10.1103/PhysRevPhysEducRes.14.020104

Yeung, S. yin S. (2015). The conception of teaching higher-order thinking: perspectives of Chinese teachers in Hong Kong. Curriculum Journal, 26(4), 553–578. https://doi.org/10.1080/09585176.2015.1053818