Pegem Journal of Education and Instruction, Vol. 12, No. 1, 2022 (pp. 292-300)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Higher-order Thinking Test of Science for College Students Using Multidimensional Item Response Theory Analysis

Ika Maryani[1*], Zuhdan Kun Prasetyo[2], Insih Wilujeng[3], Siwi Purwanti[4]

[1-3]Universitas Negeri Yogyakarta, Yogyakarta Indonesia,

[4]Universitas Ahmad Dahlan, Yogyakarta Indonesia

Abstract

The purpose of this study was to construct a higher-order thinking test of science for pre-service elementary school teachers. The test was created using the ADDIE model. The analysis stage was carried out by identifying the needs and baseline of higher-order thinking skills of students from the department of primary School Teacher education in Yogyakarta. The design stage involved the creation of test blueprints and questions cards. The development stage involved validating the test's content and construct validity. The content validity test was conducted using the Delphi technique with seven validators, whilst the construct validity test was conducted using item response theory and EFA. This study developed 77 questions, 73 multiple choice questions, and four essay questions, all of which were determined to be valid in terms of content and constructions. The HOTS test's content validity test resulted in a V-value of 0.879 (valid with high criteria) based on the average Aiken's V index. Meanwhile, reliability analysis using the Cronbach's Alpha coefficient revealed a score of 0.907 for the 77 test items based on the construct validity test. The discriminatory index (di) classified all items as good, whereas the difficulty index (bi) classified 63 items as good and 10 as poor. The ten items were revised, despite their high index of difference. All of the test questions are appropriate for students whose ability score (θ) ranged from -2.85 to 2.15.

Keywords: ADDIE, Higher-order thinking, Science, Test.

Introduction

The twenty-first century, with its ultramodern qualities, enables upheaval in several spheres of life, as well as a rapid renewal process that necessitates community preparation. The readiness of the educational environment is one of the absolutes. Education must be standardized to meet the needs of the twenty-first century. Teachers in the twenty-first century will encounter much more problems than in the previous centuries (Andriani, 2010). Teachers are confronted with a far more varied student population, more complicated and demanding subject matter, a higher quality of learning, and increased expectations for students' higher thinking abilities (Darling-Hammond, 2006). This represents a significant challenge for Institutions of Personnel Education in terms of developing teacher candidates who possess these competencies. LPTK graduates must possess strong critical thinking skills to aid in the school-based learning process. In LPTK, the stages of student learning correspond to those of adult learners (andragogy). At this level, students exhibit eight critical qualities of learning: 1) they are self-directed, 2) they are practical and goal-oriented, 3) they are more resistant to change due to their lack of openness, and 4) they learn more slowly and hence require integrative knowledge, 5) they value personal experience as a source of learning, 6) they are highly motivated, 7) they take on multiple responsibilities, and 8) they have high expectations (Pappas, 2013). Science education is one of the critical lessons that aspiring elementary school teachers at the Department of Primary School Teacher Education must know. Numerous PSTE study programs have a hierarchical structure for science courses based on their study materials and depths. In general, all science courses are designed to provide PSTE students with pedagogical and content knowledge (PCK). As a result, pre-service primary school teachers are competent to create and develop science instruction independently or in conjunction with other subjects. The characteristics of science learning are complex and need advanced analytical and critical thinking abilities, posing a variety of difficulties for students who have not mastered them. Among them include misconceptions about science (Faizah, 2016), learning difficulty in science (Maryani et al., 2018)composing, and presenting ideas. The high complexity causes many cases of learning difficulties. This study aims to diagnose the learning difficulties that occur on 5th-grade elementary school students. The research was conducted in Muhammadiyah Pakem Elementary School,

Corresponding Author e-mail: ika.maryani@pgsd.uad.ac.id

https://orcid.org/0000-0002-7154-2902

How to cite this article: Maryani I, Prasetyo ZK, 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, Vol. 12, No. 1, 2022, 292-300

Source of support: Nil

Conflict of interest: None.

DOI: 10.47750/pegegog.12.01.30

Received: 12.10.2021

Accepted: 08.12.2021 **Publication:** 01.01.2022

Sleman, Yogyakarta Special Regency. The subjects were 29 of 5th-grade elementary students. Data collection techniques were interviews, tests, and documentation. Data analysis techniques were descriptive statistic as a quantitative analysis and interactive model as a qualitative analysis. The learning difficulties were diagnosed by describing the students who were identified having learning difficulties; localizing the difficulties; and determining the factors that cause learning difficulties. The results showed that the difficulties experienced by students were in basic competence 1.1–1.5 (human blood circulation organs, and poor learning outcomes in science. Another issue that PSTE students face is the overwhelming amount of study materials that must be memorized. In this department, elementary school students must study five core subjects and additional competency support courses. These students are required to master the principle of each learning model and develop it as innovative learning in elementary schools. This objective can be met if pre-service teachers possess strong critical thinking skills and the ability to adjust to changing circumstances. This capability is encapsulated in numerous studies on 21st-century skills.

Numerous education organizations and experts have researched 21st-century skills. The Assessment & Teaching of 21st Century Skills (ATC21S) classifies 21st-century skills into four areas, one of which is a manner of thinking (Suto, 2013). Not only rich countries are monitoring the issue of 21st-century skills; Indonesia is also participating in the study. Critical thinking, problem-solving, communication, and collaboration skills are all necessary for the twenty-first century (Trisdiono, 2013). Cognitive processes establish an individual's foundation when confronted with life's issues. A cognitive process is divided into various stages, including remembering, comprehending, applying, analyzing, making a judgment, and decision making. These elements of thinking are then referred to as Higher Order Thinking Skills (HOTS)

Higher Order Thinking Skills (HOTS) assess thinking abilities that go beyond recall and memorization to include features of analysis, synthesis, and evaluation. HOTS are cognitive abilities that result in higher-level thinking (Alice Thomas & Glenda Thorne, 2009). Higher-level thinking is intended to be more than the regurgitation of information. Higher Order Thinking Skills are critical for adult learners, particularly in developing scientific concepts and applying them in everyday life, including in all university courses. In a nutshell, HOTS teach individuals how to analyze, synthesize, and evaluate (Alice Thomas & Glenda Thorne, 2009).

Research on pupils' cognitive abilities has been conducted in Indonesia. One of them demonstrates that elementary school students in Semarang, Indonesia, lack critical thinking skills. The learning process is stymied by evaluation objectives that focus only on lower-order thinking skills. Additionally, pupils' ability to categorize induced thinking is moderate. Students' capacity to deduce, analyze errors, develop an analytical perspective, make decisions, gain experience, and solve problems is rated as low (Fajriyah & Agustini, 2018). *The low thinking abilities of elementary school pupils in Indonesia are a result of a variety of circumstances, including the continuing emphasis on developing low-level thinking abilities (Surya et al., 2018). Most teachers continue to struggle with teaching and familiarizing their students with higher-order thinking. This is due to a teacher shortage of information about how to plan and administer HOTS instruction (Kuntarto et al., 2019). Similar circumstances exist for elementary school teacher candidates (pre-service teachers). According to studies (Gradini et al., 2018; Wiyoko & Aprizan, 2020), the proportion of pre-service elementary school teachers who fall into the LOTS category is greater than the proportion of pre-service elementary school teachers who fall into the HOTS category.*

Many studies have developed higher-order thinking skills (HOTS) tests of science; however, they mostly refer to Bloom's Taxonomy (Abdullah et al., 2015; Atmojo et al., 2017; Utomo et al., n.d.; Zulfiani et al., 2020)work, and be scientific and communicate it as an important aspect of Life Skills. Science learning emphasizes the provision of direct learning experience through the use and development of process skills and scientific attitudes, so as to empower the high thinking ability of Elementary School Pre-Service Teacher (ESPT, Few have examined the HOTS features of alternative theories that better fit the needs of 21st-century learning. With regards to this issue, we believe it is critical to construct a higher-order thinking skills (HOTS) test of science that relates to a variety of cognition/taxonomy theories that are tailored to the 21st century's issues.

Method

Research design

This Research and Development (R and D) study employed the ADDIE development method, which consisted of the following stages: analysis, design, Develop, Implement, and Evaluate (Branch, 2010). The research design is presented in Figure 1.

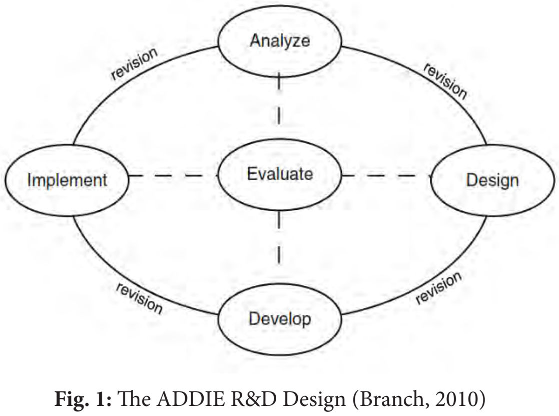

Fig. 1: The ADDIE R&D Design (Branch, 2010)

As illustrated in Figure 1, the ADDIE development design comprises five interdependent stages. At the Analyze stage, needs analysis for HOTS-based test development and problem analysis was performed. At the design stage, the product design and prototype were generated. At the development stage, product revision, content validity test, and construct validity test were carried out to ensure the validity of the final product. The implementation stage was responsible for the overall product implementation process. At each level, the product can be revised, and the process and results of product deployment can be evaluated.

Participant

The samples have been taken randomly on elementary school teacher education students in Yogyakarta. Seven experts evaluated the content of the product under development, and 268 students participated in the construct validity test.

Data Collection Tools

The HOTS test was divided down into six indicators, namely logic and reasoning, analysis, evaluation, and creation, problem-solving, and judgment. Each indicator was developed into 7-10 questions to produce 77 questions. Content validity was assessed using a questionnaire while construct validity was measured using the developed questions.

Data Analysis

The content validity test was conducted using the Delphi technique. The results of the validity test were analyzed using Aiken's V, whilst the construct validity test findings were evaluated using item response theory.

Findings

This study was successful in creating 77 HOTS test items, which included 73 multiple-choice questions and four essay questions. Validator feedback on the HOTS-based test instrument under development was just as valuable as input on other products. The validators checked the adequacy of learning achievement-learning indicators-question indicators-and items more thoroughly. The usage of analogies and experimental data was re-examined considering their logical consistency under specific settings. The editorial questions, the stimulus, the form of several items from multiple choice to description, as well as the response possibilities for multiple-choice questions, have all been altered significantly. The following summarizes the validators' input.

a. Writing

b. For test-item indicators, use the KKO analyzed from books written by Marzano or Anderson dan Krathwoll.

c. Input for the test items

i. The HOTS instrument should be re-examined to determine whether the posed questions are rational. For instance, question number one says "when throwing a baseball from a distance of 7 meters, can the bounce travel as far as 10 meters with the power of an ordinary person?"

ii. Question No. 2 is similarly less specific in terms of the ABCD points' position. Are these dots consecutive or non-sequential? Answers are frequently skewed. The solution to Problem No. 6 is ambiguous: the applicable laws are Newton's III and Pascal's laws, but Pascal's laws do not include mechanics.

iii. The illustration is unclear, as in point No. 4 regarding the top of the hill. Problems can trap students because they believe that what is anticipated is the absence of frictional force, and hence refuse to consider alternative explanations for the correct answer.

iv. Certain questions, particularly those regarding "creation", should be transformed into essay questions.

Following modifications to the HOTS instrument, it was reviewed using an assessment sheet. The HOTS instrument was evaluated on ten dimensions, including a) the items' suitability for learning outcomes; b) the items' suitability for the HOTS indicators; c) the items' suitability for the question indicator; d) the stimulus' novelty (encouraging students to read); e) the stimulus' quality (contextual and implies the answer to the question); f) the suitability of the item with the material being learned; g) the ability to measure HOTS in aspects of logic, reasoning, analysis, evaluation, creation, problem-solving, and judgment; h) clarity of the formulation of the questions; i) clarity and arrangement of answer choices on multiple-choice questions (homogeneous) and j) use of language. Additionally, the HOTS instrument makes use of the Likert scale. The instrument's content validity test indicated that the average Aiken V index produced V = 0.879 (highly valid). As a result of expert validation, the HOTS instrument was determined to be valid and was used in the next stage, namely the construct validity test.

The Results of the Construct Validity Test on the HOTS Instrument

a) Test of Unidimensionality Assumption

The criterion for meeting this assumption is that each test item evaluates only one ability. The assumption can be tested using factor analysis, which generates KMO, eigenvalues, explainable

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.830
Bartlett's Test of Sphericity	Approx. Chi-Square	6260.265
	df	2926
	Sig.	.000

variance, and factor components. SPSS 24 was used to conduct the exploratory factor analysis. The following summarizes the findings of the factor analysis.

The outcome of factor analysis indicates that the KMO value is 0.830 or greater than 0.50, indicating that the sample size utilized in this trial is adequate. Additionally, the Bartlett test's chi-square value is 6260.265 with 2926 degrees of freedom and a p-value greater than 0.01. Intercorrelation between variables was determined using the KMO-MSA test (Widarjono, 2015). If the matrix has a KMO value greater than 0.5, it can be factored in.

A test is considered unidimensional if it is demonstrated to measure only one dominant dimension, namely homogenous ability (Widarjono, 2015). The number of factors created can be determined by the presence of eigenvalues greater than one, which is the indicator factor (Widarjono, 2015). Factor analysis (Appendix 1) identified 27 components with an eigenvalue greater than one. This indicated that the 73 HOTS test items comprise 13 factors. The analysis results indicated that factor 1 is the dominant factor due to its eigenvalue of 12.931, which is greater than the others or the most dominant, implying that the HOTS test is unidimensional.

Statistical analysis also indicated an eigenvalue of 12.931, where the result is more than 2 times the eigenvalue of the second factor with a percentage of the variance of 16.79%. Cumulatively, the percentage of the 27 factors is 65.546, suggesting that 65.546% is explained by the 27 existing components. The cumulative percentage of 65.546% has fulfilled the minimal condition for the cumulative value of taking the proper number of variables, which is 50% (Widarjono, 2015). Evidence of cumulative percentage values corroborates the notion that the HOTS test instrument is believed to be unidimensional.

Dimensions recorded in data can be proven in the scree plot findings, specifically the number of steeps. The number of steps shows the number of dimensions/factors, while the slope of the change in eigenvalues does not indicate the presence of dimensions (Widarjono, 2015). Therefore, unidimensionality can also be shown from the ensuing scree plot. The test is deemed to be unidimensional when components 1 and 2 in the scree plot have a high enough distance (Furr & Bacharach, 2008).

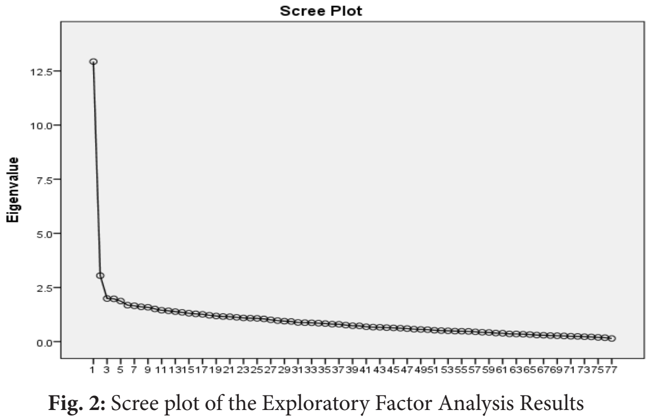

Fig. 2: Scree plot of the Exploratory Factor Analysis Results

According to the scree plot in Figure 2, component 1 is located far away from component 2, whereas component 2 is located quite close to component 3 and other components. Additionally, as illustrated in Figure 2, the eigenvalues begin to slope with the third component. This demonstrates a single dominant factor and that other factors contribute significantly to the variance explained. The unidimensional analysis results corroborate the assertion of (Widarjono, 2015) that this HOTS test evaluates at least two components, with the first factor serving as the dominant factor. The scree plot in Figure 2 demonstrates that the HOTS test currently under development is unidimensional.

b) Test of Local Independence Assumption

One of the conditions for IRT analysis is the assumption of local independence. This assumption test is used to determine whether students' abilities are independent of the test questions, which means that their responses to one item do not affect their responses to subsequent items. The unidimensionality of the student response data to the test automatically establishes the local independence assumption test (Widarjono, 2015). The local independence assumption, on the other hand, can be demonstrated using a covariance matrix based on the ability of pupils categorized into many groups. If the correlation between the capability intervals is modest or close to zero, this assumption is fulfilled. Thus, a covariance value near zero satisfies the local independence assumption. Table 1 contains the covariance matrix.

Table 1 presents the variance-covariance matrix values for several groups of students' skills. The analysis reveals that the covariance variation across groups of students' ability intervals that form a diagonal line is negligible if not nil. As there is no association between the two variables, the assumption of local independence is satisfied.

c) Test of Parameter Invariance Assumption

The third requirement is parameter invariance. Parameter invariance shows that the test items are independent of the distribution of the students' ability parameter and vice versa, that students' ability parameter is independent of the test items. Students' abilities will not change because of working on a package of questions with distinct item parameters, and the item parameters will remain constant regardless of which group of students is assessed. There are two types of parameter invariance. The first type is item parameter invariance, and the second type is ability parameter invariance. The invariance of the item parameter can be determined by dividing the sample (218 students) into two even and odd groups.

Table 1: Covariance Matrix of Students' Higher Order Thinking Skills (HOTS)[c]

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0,0726									
K2	0,0227	0,0132								
K3	0,0130	0,0066	0,0036							
K4	0,0250	0,0090	0,0052	0,0098						
K5	0,0077	0,0036	0,0020	0,0031	0,0012					
K6	0,0062	0,0024	0,0014	0,0023	0,0008	0,0006				
K7	0,0233	0,0089	0,0050	0,0089	0,0029	0,0022	0,0083			
K8	0,0092	0,0044	0,0023	0,0034	0,0013	0,0009	0,0033	0,0016		
K9	0,0312	0,0124	0,0070	0,0122	0,0040	0,0030	0,0113	0,0046	0,0156	
K10	0,0813	0,0555	0,0251	0,0300	0,0129	0,0095	0,0327	0,0163	0,0442	0,7280

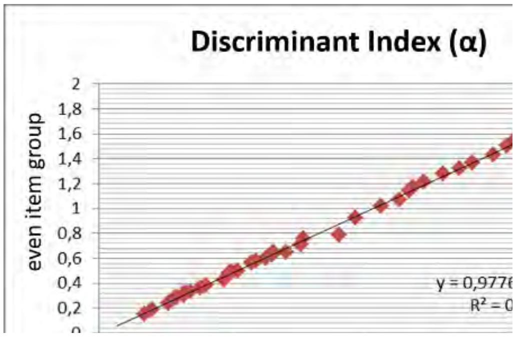

Fig. 3: Scree plot of the parameter invariance of the HOTS test's discriminatory power

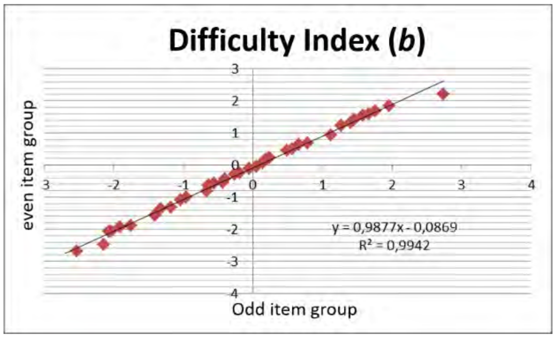

Fig. 4: Scree plot of the Parameter Invariance of HOTS Test Difficulty Level

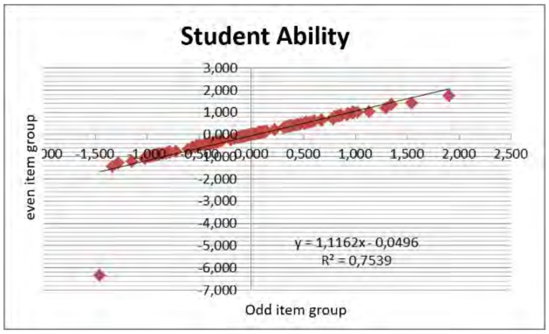

Fig. 5: Scree plot of the Parameter Invariance of Students' Ability

The estimated grain parameters for each sample are then plotted and associated using a scree plot. If the correlation is positive and significant, the assumption of item parameter invariance is satisfied (Widarjono, 2015). Figure 3 illustrates the estimation results for the invariance of the item parameters.

The scree plot in Figure 3 depicts the estimation of item parameter invariance for item discriminating power after students worked on odd and even questions. As illustrated in Figure 3, the estimated values are spread out and reasonably close to the linear line. The discriminatory power has a strong correlation with the student's response to the odd and even test items (0.9962). The scree plot and correlation analysis indicate that the discriminating power of the test items is invariant. The separation of two groups of test takers, odd and even, was also incorporated in the item parameter analysis for difficulty level. Microsoft Excel was used to conduct the analysis. Figure 4 depicts the correlation between the findings of the analysis.

The scree plot of the estimated invariance of items in terms of difficulty level after students worked on odd and even questions is shown in Figure 4. As illustrated in Figure 4, the estimated values are dispersed and somewhat close to the linear line. The correlation coefficient between the difficulty of the questions and the responses of students to odd and even items is 0.9942 (high). Figures 3 and 4 indicate that the estimation of the item parameters' invariance in terms of discriminatory power and difficulty level is satisfactory.

After splitting the odd and even subtest groups, the invariance of the students' ability parameter can be examined. The estimated ability parameter for each sample was plotted and associated using a scree plot. If the correlation is significant and positive, the assumption of invariance of the student's ability parameters is satisfied (Widarjono, 2015). In general, students' capacity to work on the test is estimated as scattered (Figure 5).

The scree plot in Figure 5 depicts the estimated invariance of students' ability following an analysis of the abilities of even and odd-numbered students. Additionally, the scree plot findings show why the estimated values are quite close to a straight line with a correlation coefficient of 0.7539 (very high). In conclusion, the ability parameter invariance assumption has been satisfied.

d) Estimation of Reliability

The reliability coefficient of an instrument indicates the degree of confidence in the error-free findings of measurement (the greater the reliability coefficient, the more accurate the measurements). In this study, reliability was estimated using the SPSS 24 program. The Cronbach's Alpha coefficient for 77 items was 0.907 based on confirmed data. According to Mahrens and Lehman, while there is no universal agreement, it is usually believed that the test used to make individual student placement decisions must have a minimum reliability coefficient of 0.85 (Mehrens & Lehmann, 1991). According to the findings of this study and the experts' view, the reliability of the test developed in this study meets the criteria for a reliable test.

e) Model Fit

The three assumptions for the IRT analysis had been well fulfilled so that the HOTS multiple-choice test was examined for model fit. Seventy-three items were produced. The model fit test for 1-PL, 2-PL, or 3-PL was performed by comparing X^2. The probability value for each test item must fulfill $p > 0.05$. The model fit analysis results summarized in Appendix 1 indicate that the 2-PL model is the best appropriate for the HOTS test instrument. In comparison to the 1-PL or 3-PL models, the 2-PL model accommodates the majority of the HOTS test's multiple-choice items. Since the study requires a 2-PL model, the parameters to examine are the discriminatory power (a) and the difficulty level (b) of each test item. Items that do not match the criteria for a "good item" are omitted from the final product.

Table 3: Model Fit Test on HOTS Essay Questions

	X2				
Item	*Statistics*	*df*	*RMSEA*	*P-Value*	*Remarks*
A26	0.581	4	0.000	0.965	Fit
A8	3.771	5	0.000	0.583	Fit
A36	7.614	3	0.076	0.055	Fit
A15	4.749	4	0.026	0.314	Fit

After examining the model fit on the -choice test, the HOTS test's four essay items were analyzed. For essay questions, the model fit criteria are identical to those for multiple-choice questions. The essay questions, on the other hand, were examined using the R package MIRT program. This was done because the essay questions were scored as polytomous, which prevented them from being examined using the BILOG-MG tool. TABLE 3 summarizes the model fit analysis of the HOTS test essay questions.

According to TABLE 3, all test items fit the 2-PL model applied. The examination of the multiple-choice items and essay questions reveals that the 2-PL model is the best fit for the HOTS test items. The parameters measured in both types of questions are the same, namely discriminatory power (a) and degree of difficulty (b) of each test item.

f) Parameter of Time Item

The 2-PL model was used to determine the characteristics of a good test item. The test items that fit the 2PL model were re-analyzed to determine their properties. According to the 2-PL model, the requirements for a good item are based on the discriminatory power (ai) and level of difficulty (bi) of each item. Discriminating power is regarded to be good if it is between 0 and 2. Additionally, a good difficulty index should range between -2 and +2 (Widarjono, 2015). This study found the discriminant index and the difficulties index of 73 questions (Appendix 2). These findings indicate that all items have a high discriminatory power index (ai), while 63 test items have a good difficulty level (bi) and ten items have a low difficulty level (bi). Although the 10 items showed a high discrimination index, they had a low difficulty level. Therefore, the ten items (A29, A30, B14, A27, A33, B25, A28, A17, A25, and B19) were revised.

The analysis of the multiple-choice test parameters was then continued with the analysis of the HOTS essay questions. The essay questions were analyzed using the R-Program. The results of the parameter analysis of the essay questions are shown as follows.

As shown in Table 4, item A26 has a low discrimination index of 7.717. Nevertheless, items A8, A46, and A15 have high discriminatory indices. All essay items have a reasonable difficulty index. Based on these findings, item A26 has a low discriminatory index but a high difficulty index; hence, item

Table 4: The Results of Parameter Analysis on the HOTS Essay Questions

	Discriminatory Power		*Difficulty Level*				
Item	*a*	*Remarks*	*b*	*b2*	*b mean*	*Remarks*	*Conclusion*
A26	7.717	Poor	-0.981	-0.130	-0.555	Good	Revised
A8	0.07	Good	-0.851	-0.434	-0.642	Good	Accepted
A46	1.402	Good	-0.865	1.871	0.503	Good	Accepted
A15	0.173	Good	-0.260	-	-0.260	Good	Accepted

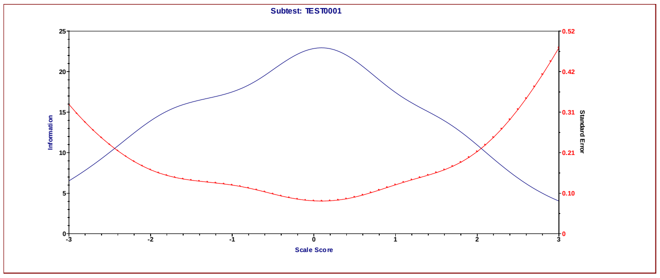

Fig. 6: IF & SEM Curves

A26 must be amended and items A8, A46, and A15 were accepted.

g) Information Function and Standard error of measurement (IF SEM &)

The test information function is equal to the sum of the test item functions. The relationship between the test information function and the standard error of measurement (SEM) is inverse, with a higher test information function indicating a smaller measurement error and vice versa. Figure 6 illustrates the IF and SEM curves.

The analysis of the 2-Parameter Logistics (2-PL) model using BILOG-MG yields discriminating power (ai) and item difficulty level (bi), which were then utilized to determine the information function value for each HOTS test item. The test information value was calculated by adding the information functions of each item. The maximum test information function is found in ability = 0.1, with a value of 23.2 and a measurement error of 0.7. Additionally, Figure 6 illustrates that the HOTS test instrument covers the interval's lower and higher bounds. The interval's lower and upper bounds are the ability scores at which the graphs of the information function and standard error of measurement overlap. Based on the intersection line, it was determined that the HOTS test established in this study is appropriate for assessing higher-order thinking skills in students with an ability (θ) of -2.85 to 2.15.

Discussion

Higher-order thinking skills (HOTS) are higher-level cognitive abilities, not only memorization. HOTS entail several mental processes, including analyzing, evaluating, and producing, all of which are embedded in the problem-solving process. According to (Lewy, 2011), any ability that requires analysis, evaluation, and production is classified as a higher-order thinking skill. Bloom's Taxonomy is the most frequently accepted hierarchical arrangement of HOTS in the field of education, as it examines the levels of thinking from knowledge to evaluation (Ramos et al., 2013). However, the new paradigm of educational research frequently references Marzano's Taxonomy, which includes comparing, classifying, inductive reasoning, deductive reasoning, error analysis, construction support, perspective analysis, abstracting, decision making, investigation, problem-solving, experimental inquiry, and invention (Heong et al., 2011, 2016; Marzano, 1993; Marzano & Kendall, 2006).

According to Marzano's Taxonomy, higher-order thinking skills (HOTS) enable the development of student learning outcomes (SLO), class activities, and learning performance (Dubas & Toledo, 2016; Toledo & Dubas, 2016). Students that possess higher-order thinking skills are capable of learning, improving their performance, and overcoming their weaknesses (Yee et al., 2011). Students who received thinking skills training improved their reading comprehension and academic performance. This demonstrates the critical nature of thinking skills in resolving learning challenges, stimulating competitive thinking, creating intellectuals, and avoiding cognitive errors (Heong et al., 2011). Higher-order thinking skills are classified according to a level of cognition (cognitive capacity). The most often used classification of thinking abilities is Bloom's Taxonomy or its modification, which includes the following: 1) remembering, 2) comprehending, 3) applying, 4) analyzing, 5) evaluating, and 6) creating (C. A. Anderson & Krathwohl, 2014; L. W. Anderson et al., 2000). Numerous scholars classify HOTS into three categories: analysis, evaluation, and creation.

Marzano defines knowledge as "information, mental procedures, and psychomotor procedures." Following that, the domain is separated into six hierarchical cognitive processes: retrieval, comprehension, analysis, knowledge utilization, metacognition, and self-system thinking. Marzano defines HOTS as the following: comparing, classifying, inductive reasoning, deductive reasoning, error analysis, construction support, perspective analysis, abstracting, decision-making, investigation, problem-solving, experimental inquiry, and invention (Heong et al., 2011, 2016; Marzano, 1993; Marzano & Kendall, 2006).

Along with Bloom, Anderson, and Marzano, Webb (2002) provides stages of thinking that are commonly employed in standard measurement in many nations. This thinking stage consists of four levels, namely 1) recall and reproduction, 2) skills and concepts, 3) strategic thinking, and 4) extended thinking. The SOLO (Structure of Observed Learning Outcomes) Taxonomy is another cognitive taxonomy that is commonly used in Australia, New Zealand, Canada, and the United Kingdom. SOLO is a hierarchical taxonomy of cognitive abilities that focuses on distinct elements and their relationships. This hierarchy is divided into five levels: pre-structural, uni structural, multi-structural, relational, and extended abstract. Brookhart (2010) constructs HOTS indicators using slices from all four taxonomies. According to Brookhart (2010), HOTS consist of logical ability and reasoning, analysis, evaluation, and creation, problem-solving, judgment, and creativity, and creative thinking. However, this study used a taxonomy that is tailored to the demands of future primary school science instructors.

Conclusion

This study developed 77 questions, 73 multiple choice questions, and four essay questions, all of which were determined to be valid in terms of content and constructs. The content validity test, calculated using the average Aiken V index, produced V = 0.879, indicating that the HOTS test is highly valid. Cronbach's Alpha coefficient for 77 items is 0.907 based on the construct validity test. The analysis of the multiple-choice items and essay questions revealed that the 2PL model was the most appropriate form of IRT model for analyzing the test items. Each HOTS test item has a discriminatory power index (ai) in the good category. However, in terms of difficulty level index (bi), there were 63 items in the good category and 10 items in the bad category. As a result, the ten test items needed to be altered. The 10 items had a high discriminatory index but a low level of difficulty. Items that needed to be revised included A29, A30, B14, A27, A33, B25, A28, A17, A25, and B19. Item A26 in the essay question section showed a low discrimination index, but a high difficulty level. Therefore, item A26 was revised, but items A8, A46, and A15 were accepted. All the test questions generated in this study are appropriate for assessing the higher-order thinking skills of students with the ability (θ) ranging from -2.85 to 2.15.

Suggestion

This study recommends further research to be able to promote HOTS through a learning approach. This aims to increase the HOTS of students throughout Indonesia.

Limitation

This research is limited only to the development of Physics Science test instruments. This is based on a pre-research needs analysis. Development in other fields is recommended.

References

Abdullah, A. H., Ali, M., Liyana, N., & Abidin, Z. (2015). Analysis of students' errors in solving higher-order thinking skills (HOTS) problems for the topic of a fraction. *Asian Social Science*, *11*(21), 133–142. https://doi.org/10.5539/ass.v11n21p133

Alice Thomas, & Glenda Thorne. (2009). *How to Increase Higher Level Thinking | Center for Development and Learning.* The Center for Development And Learning. http://www.cdl.org/articles/how-to-increase-high-order-thinking/

Anderson, C. A., & Krathwohl, D. R. (2014). Bloom's Taxonomy of Educational Objectives. *Center for Innovation in Teaching and Learning*, 1–2.

Anderson, L. W., Krathwohl, D. R., & Airasian, P. W. (2000). *A Taxonomy for learning, teaching, and assessing: a revision of bloom's taxonomy of educational objectives.* http://digilib.uns.ac.id/pengguna.php?mn=showview&id=21967

Andriani, D. E. (2010). Mengembangkan Profesionalitas Guru Abad-21 Melalui Program Pembimbingan yang Efektif. *Jurnal Manajemen Pendidikan*, *6*(2), 78–92. http://journal.uny.ac.id/index.php/jmp/article/view/3639%7B%25%7D0Ahttp://journal.uny.ac.id/index.php/jmp/article/download/3639/3112

Atmojo, I., Sajidan, P., Sunarno, W., & Ashadi, M. (2017). Profile of elementary school pre-service teacher based on high order thinking skills (HOTS) on natural science subject. *Advances in Social Science, Education, and Humanities Research (ASSEHR)*, *158*, 501–504. https://doi.org/10.2991/ICTTE-17.2017.57

Branch, R. M. (2010). *Instructional Design: The ADDIE Approach*. Springer US. https://doi.org/10.1007/978-0-387-09506-6

Brookhart, S. M., Bookhart, S. M., & Brookhart, S. M. (2010). *How to Assess Higher-Order Thinking Skills in Your Classroom*. ASCD. http://www.ascd.org/publications/books/109111.aspx

Darling-Hammond, L. (2006). Constructing 21st-Century Teacher Education. *Journal of Teacher Education*, *57*(3), 300–314. https://doi.org/10.1177/0022487105285962

Dubas, J. M., & Toledo, S. A. (2016). Taking higher-order thinking seriously: using Marzano's taxonomy in the economics classroom. *International Review of Economics Education*, *21*, 12–20. https://doi.org/10.1016/j.iree.2015.10.005

Faizah, K. (2016). Miskonsepsi dalam pembelajaran IPA. *Darussalam: Jurnal Pendidikan Komunikasi Dan Pemikiran Hukum Islam*, *8*(1), 113–125.

Fajriyah, K., & Agustini, F. (2018). Analisis Keterampilan BerpikirTingkat Tinggi Siswa SD Pilot ProjectKurikulum 2013 Kota Semarang. *Elementary School: Jurnal Pendidikan Dan Pembelajaran Ke-SD-An*, *5*(1), 54–67. https://doi.org/10.31316/esjurnal.v5i1

Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics. An introduction*. Sage.

Gradini, E., Firmansyah, F., & Noviani, J. (2018). Menakar kemampuan berpikir tingkat tinggi calon guru matematika melalui level HOTS Marzano. *Eduma : Mathematics Education Learning and Teaching*, *7*(2), 41–48. https://doi.org/10.24235/eduma.v7i2.3357

Heong, Y. M., Othman, W. B., Yunos, J. B. M., Kiong, T. T., Hassan, R. Bin, & Mohamad, M. M. B. (2011). The Level of Marzano Higher Order Thinking Skills among Technical Education Students. *International Journal of Social Science and Humanity*, *1*(2), 121–125. https://doi.org/10.7763/ijssh.2011.v1.20

Heong, Y. M., Sern, L. C., Kiong, T. T., & Binti Mohamad, M. M. (2016). The Role of Higher Order Thinking Skills in Green Skill Development. *MATEC Web of Conferences*, *70*, 1–5. https://doi.org/10.1051/matecconf/20167005001

Kuntarto, E., Alirmansyah, A., & Kurniawan, A. R. (2019). Kemampuan mahasiswa PGSD dalam merancang dan melaksanakan pembelajaran berbasis high order of thinking skills. *Jurnal Kiprah*, *7*(2), 107–116. https://doi.org/10.31629/KIPRAH.V7I2.1454

Lewy. (2011). Pengembangan Soal Untuk Mengukur Kemampuan Berpikir Tingkat Tinggi Pokok Bahasan Barisan dan Deret Bilangan di Kelas IX Akselerasi SMP Xaverius Maria Palembang. *Jurnal Pendidikan Matematika*, *5*(1), 58–73. https://doi.org/10.22342/jpm.5.1.821

Maryani, I., Husna, N. N., Wangid, M. N., Mustadi, A., & Vahechart, R. (2018). Learning difficulties of the 5thgrade elementary school students in learning human and animal body organs. *Jurnal Pendidikan IPA Indonesia*, *7*(1). https://doi.org/10.15294/jpii.v7i1.11269

Marzano, R. J. (1993). How classroom teachers approach the teaching of thinking. *Theory Into Practice*, *32*(3), 154–160. https://doi.org/10.1080/00405849309543591

Marzano, R. J., & Kendall, J. S. (2006). *The new taxonomy of educational objectives* (2nd ed.). Corwin Press.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology.* Harcourt Brace College Publishers.

Pappas, C. (2013). *8 important characteristics of adult learners.* Elearning Design and Development. https://elearningindustry.com/8-important-characteristics-of-adult-learners

Ramos, J. L. S., Dolipas, B. B., & Villamor, B. B. (2013). Higher Order Thinking Skills and Academic Performance in Physics of College Students : A Regression Analysis. *International Journal of Innovative Interdisciplinary Research*, *4*(1), 48–60. https://doi.org/ISSN 1839-9053

Surya, A., Sularmi, S., Istiyati, S., & Prakoso, R. F. (2018). Finding HOTS-base mathematical learning in elementary school students. *Social, Humanities, and Educational Studies (SHEs): Conference Series*, *1*(1), 30–37. https://doi.org/10.20961/shes.v1i1.24308

Suto, I. (2013). *21 st Century skills : Ancient, ubiquitous, enigmatic ?*

Toledo, S., & Dubas, J. M. (2016). Encouraging Higher-Order Thinking in General Chemistry by Scaffolding Student Learning Using Marzano's Taxonomy. *Journal of Chemical Education*, *93*(1), 64–69. https://doi.org/10.1021/acs.jchemed.5b00184

Trisdiono, H. (2013). *Strategi pembelajaran abad 21.*

Utomo, A. P., Narulita, E., & Shimizu, K. (n.d.). *Diversification of reasoning science test items of TIMSS grade 8 based on higher-order thinking skills: a case study of Indonesian students.*

Webb, C. (2002). *The Motivational Enhancement Therapy and Cognitive Behavioral Therapy Supplement: 7 Sessions of Cognitive Behavioral Therapy for Adolescent Cannabis Users.* US Department of Health and Human Services, Substance Abuse and Mental Health.

Widarjono, A. (2015). *Analisis Multivariat Terapan dengan Program SPSS, AMOS, dan SMARTPLS, II.* UPP STIM YKPN.

Wiyoko, T., & Aprizan, A. (2020). Analisis profil kemampuan kognitif mahasiswa PGSD pada mata kuliah ilmu alamiah dasar. *IJIS Edu : Indonesian Journal of Integrated Science Education*, *2*(1), 2020. https://doi.org/10.29300/ijisedu.v2i1.2384

Yee, M. H., Othman, W., Md Yunos, J., Tee, T. K., Hassan, R., & Mohamad, M. M. (2011). The level of Marzano higher-order thinking skills among technical education students. *International Journal of Social Science and Humanity.* http://merr.utm.my/id/eprint/1589

Yusuf, I., Widyaningsih, W., & Sebayang, R. B. (2018). Implementation of e-learning based-STEM on quantum physics subject to student HOTS ability. *Journal of Turkish Science Education*, *15*(Special), 67–75. https://doi.org/10.12973/tused.10258a

Zulfiani, Z., Suwarna, I. P., & Sumantri, M. F. (2020). Science adaptive assessment tool: Kolb's learning style profile and student's higher-order thinking skill level. *Jurnal Pendidikan IPA Indonesia*, *9*(2), 194–207. https://doi.org/10.15294/JPII.V9I2.23840