Volume: 8
Nomor : 4
Bulan : November
Tahun : 2022

E-ISSN: 2656-940X
P-ISSN: 2442-367X
URL: jurnal.ideaspublishing.co.id

Analysis of HOTS Instrument for Prospective Physics Teacher Using Generalized Partial Credit Model

Duden Saepuzaman
[1)]Universitas Negeri Yogyakarta
[2)]Universitas Pendidikan Indonesia

Edi Istiyono
Haryanto
Universitas Negeri Yogyakarta

Pos-el: dsaepuzaman@upi.edu
edi_istiyono@uny.ac.id
haryanto@uny.ac.id

DOI: 10.32884/ideas.v8i4.976

Abstract
This study aims to analyze item parameters characteristics and estimate prospective physics teachers ability on the Higher Order Thinking Skills (HOTS) Instrument using the Generalized Partial Credit Model (GPCM). The research subjects were 251 prospective physics teacher students in two universities concerned with producing prospective teacher graduates. Two-Tier Multiple Choice (TTMC) forms with a polytomous score of four categories make up the test instrument. Data analysis includes two stages: testing assumptions and the suitability of the Politomus IRT model. The results showed that the most suitable polytomous scoring IRT model was GPCM2PL. The item parameter analysis for the HOTS Instrument test shows the discriminating power parameter (a) value for all items included in the good category, namely the interval 0.00 to 2.00. The difficulty level analysis (b) also shows the percentage of 100% of items included in the medium category because the b value of all items is in the interval (-2) to (2).
Keywords
HOTS, prospective physics teacher, generalized partial credit model

Abstrak
Penelitian ini dimaksudkan untuk menganalisis karakteristik parameter butir dan mengestimasi kemampuan calon guru fisika pada instrumen tes Higher Order Thinking Skills (HOTS) dengan menggunakan Generalized Partial Credit Model (GPCM). Sumber data berasal dari respon calon guru fisika pada instrumen HOTS. Subyek penelitian ini adalah 251 calon siswa guru fisika di dua LPTK (LPTK merupakan perguruan tinggi yang menghasilkan calon guru). Instrumen tes berbentuk Two-Tier Multiple Choice (TTMC) dengan skor politomus 4 kategori. Analisis data meliputi dua tahap, yaitu pengujian asumsi Item Response Theory (IRT) dan kesesuaian model IRT Politomus. Hasil penelitian menunjukkan bahwa model IRT polytomous yang paling sesuai adalah GPCM dua parameter logistik (GPCM2PL). Analisis parameter butir instrument HOTS menunjukkan nilai parameter daya pembeda (a) untuk semua butir soal termasuk dalam kategori baik yaitu interval 0,00 sampai dengan 2,00. Analisis tingkat kesukaran (b) juga menunjukkan 100% butir termasuk dalam kategori sedang karena nilai b seluruh butir soal berada pada interval (-2) sampai (2).
Kata Kunci
HOTS, calon guru fisika, generalized partial credit model

Introduction

Education is an important aspect of the development of a country (Hasudungan & Kurniawan, 2018). Operationally, the implementation of education is regulated by using the curriculum. The curriculum is a reference for implementing education at all levels, from basic and secondary education to higher education (Hadi et al., 2018). Implementing education in higher education refers to the higher education curriculum with graduate competency standards based on the Kerangka Kualifikasi Nasional Indonesia (KKNI). By

juxtaposing, equating, and integrating the disciplines of education, on-the-job training, and work experience, the KKNI competence qualification tiering framework may offer acknowledgement of work competencies that adhere to the work structure in diverse industries. Regarding Indonesia s national education and training system, KKNI represents the nation s quality and identity (Republik Indonesia, 2012). These graduate competencies are described as learning outcomes or Learning Outcomes (LO). One of the learning outcomes in the physics education study program, for instance, is Higher Order Thinking Skills (HOTS). Everyone must have HOTS to compete and compete in the 21st Century (Hasan et al., 2022). Not surprisingly, the development of HOTS has become very important to be applied in the higher education curriculum (Behar-Horenstein & Niu, 2011; Satriawan et al., 2020). A person will have the opportunity to learn (learning) by having HOTS, be critical in giving reasons (reasoning), think creatively (think creatively), be able to make decisions (decision making), and solve problems (problem-solving) (Robinson, 2000). Furthermore, HOTS is one of the necessary aspects of the world of work (Cotton, 1997; Robinson, 2000). Based on these data, it is clear that higher education graduates are required to have HOTS to improve their performance and readiness, ability, and skills required by work (employability skills). Given the importance of the role of HOTS, it is very rational if HOTS becomes an important part of the competence or learning outcomes of higher education graduates, including in the Physics Education study program.

Training HOTS is closely related to the study of physics (Adeyemo, 2010). It is because HOTS significantly affects prospective physics teacher students performance and contributes to the learning outcomes of physics education (Ramos et al., 2013). Students will gain from HOTS in several ways, including improved concept comprehension, increased sensitivity to difficulties, problem-solving skills, and the ability to apply concepts in various contexts (Marlina et al., 2018; Rosmiati and Satriawan, 2019). In addition to directly connecting to performance, HOTS is a component of 21st-century skills (Ab Kadir, 2017; Ahrari et al., 2016; Dwyer et al., 2014; Guo & Woulfin, 2016; Saprudin et al., 2019). According to data from the Association of American Colleges and Universities from 2011, 95% of campus academic authorities agree that HOTS is a crucial skill for college graduates (S. Liu et al., 2018). This condition is not much different from that in Indonesia, which includes HOTS as part of learning outcomes in the higher education curriculum.

Some experts associate the definition of HOTS with various skills. Based on the views of several experts, mental abilities, including critical thinking and creative thinking, might be classified as HOTS (Conklin, 2012; King et al., 2010; Krulik & Rudnick, 1999; Presseisen, 1988), decision making (Presseisen, 1988) and logical, reflective, and metacognitive thinking (King et al., 2010), problem-solving problems (Brookhart, 2010; Presseisen, 1988). Talking about learning outcomes and goals in the world of education usually refers to the taxonomy of learning objectives. One of the most famous taxonomies is Bloom s taxonomy which Benjamin S. Bloom proposed in 1956 (Bloom, 1972). In Bloom s revised taxonomy, the HOTS feature is characterized by thinking that involves analysis, synthesis, and creation (Anderson & Krathwohl, 2001). The HOTS used in this study refers to the revised Bloom s taxonomy HOTS. The Revised Bloom s Taxonomy is the most basic and essential multilevel instructional system of psychological processes intended to structure appropriate thinking skills (Prakash & Litoriya, 2022). In addition, it is very in line with the learning achievement of the physics education program.

The assessment determines how well prospective physics teachers performed on the HOTS. Information regarding learning goals or accomplishments is gathered through assessment (Kizlik, 2012). In addition to knowing the achievement of learning outcomes/goals, assessment is also useful in providing an overview of the quality of the learning process (Bano et al., 2022). The use of high-quality instruments is one of the essential aspects of evaluating students learning outcomes. 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, Lee, and Linn, 2011; Xiao et al., 2018). Assume

E-ISSN: 2656-940X
P-ISSN: 2442-367X
URL: jurnal.ideaspublishing.co.id

that adding justifications to the second level of the two-tier choice question format will help test-takers use higher-order thinking abilities and justifications (Cullinane & Liston, 2011). Preparing test-takers for higher-order thinking tasks must consider the justifications that correspond to the answer options before selecting their responses. 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).

Apart from the form of the test instrument, ensuring that the evaluation results appropriately reflect students ability is another factor that must be considered. An evaluation is considered accurate if the results show the least amount of mistakes or errors possible. The test instrument s quality must be valid, trustworthy, and have excellent item parameters to produce results that correctly reflect students abilities. Item response theory and classical test theory are two methods that may be used to estimate item parameters for this purpose. Traditional test theory is said to have weaknesses. The primary weakness of traditional test theory is the inability to distinguish between the examinee and test characteristics, which can only be understood in a different context (Hambleton et al., 1991). In other words, the examination establishes the examinee s aptitude. When the test is challenging, the test-taker will do poorly. The individual will have a greater skill level if the test is simple. In other words, the subject/test taker and item characteristics are significantly correlated. Both the examinees traits and the item s attributes will vary as the examinees themselves change. Because the assessment outcomes depend on the test taker s subjects, conventional test theory cannot be applied in this case.

Item response theory contains the idea of releasing the connection between items and samples or test takers, which is a way to address the flaws in traditional test theory. Even if they work on items with various features, the examinees traits or abilities won t change. On the other hand, even if examinees execute things to the best of their skills, the objects qualities will stay the same. The item response concept is no longer based on test kits but actual items. Item response theory is based on two theories: (a) a collection of qualities, latent traits, or skills can predict (or explain) test takers performance on test items; and (b) as ability improves, the respondent s likelihood of responding an item correctly also rises. 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 Partial Credit Model (PCM), Graded Response Model (GRM), and Generalized Partial Credit Model (GPCM). This study uses GPCM analysis. The 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.

PCM analysis is widely used as an alternative to polytomous data analysis. PCM is used as an analysis that aims to analyze students critical thinking skills (Asysyifa et al., 2019). The results of this study indicate that all items are categorized as good. Further analysis related to the parameter Estimation of students critical thinking skills showed that there were no students who had the highest score on critical thinking skills, 1.67% of students had high critical thinking skills, 60% of students had average critical thinking skills, 1.67% students have low critical thinking skills, and 3.33% of students have the lowest critical thinking skills. Another research related to the use of PCM analysis was conducted by Istiyono (Istiyono, 2017). The purpose of this study was to describe the results of measuring higher-order thinking skills in physics (PhysHOTS). The results showed that each item in the instrument used was valid to measure students higher-order thinking skills ranging from very low, low, medium, high, and very high categories. Both of these studies focused on PCM analysis, which only analyzed one item parameter, namely the item difficulty level. In fact, the characteristics of an instrument are not only represented by the quality with the level of difficulty but also the index of differentiating power. The discriminatory index is very important to ensure that the items developed are able to compare high and low-ability students. Based on the background, this study focused on Item Parameter Analysis and Estimation of The Ability of Prospective Physics Teachers on The HOTS Instrument

E-ISSN: 2656-940X
P-ISSN: 2442-367X
URL: jurnal.ideaspublishing.co.id
Volume: 8
Nomor : 4
Bulan : November
Tahun : 2022

Dynamics Concept using The GPCM. The choice of dynamics material, because the dynamics material in which Newton s laws are very important in physics because it is a prerequisite for understanding other physics concepts.

Method

Research Design

This study uses a quantitative method and is a descriptive approach. This study describes quantitative data based on data analysis derived from student responses to the HOTS instrument test.

Sample and Data Collection

The research subjects were 251 students in detail, 136 students from Universitas Pendidikan Indonesia (West Java Province), and 115 from Universitas Sultan Ageung Tirtayasa (Banten Province). Two-Tier Multiple Choice (TTMC) forms with a polytomous score of four categories considered the test instrument (see Apendix) The scoring criteria used are presented in the table (Istiyono, 2019; Istiyono et al., 2020). The items in the instrument are 16 items. The test instrument is a HOTS instrument for dynamics in the form of TTMC. This instrument modifies the standard test Force Concept Inventory (FCI).

The analysis was carried out in two stages. They are first testing the unidimensional assumption. The unidimensional assumption is met if each item only assesses one ability (Retnawati, 2014). However, this assumption is difficult to fulfill strictly because of the cognitive, personality, and test-taking factors such as motivation, anxiety, and tendency to guess. Thus, the assumption of unidimensionality can be demonstrated if the test contains only one dominant component that measures the subject s achievement. In practice, the unidimensional assumption can be made from the eigenvalue plot, which shows one dominant component (Retnawati, 2014). Another way to prove unidimensionality is to look at the percentage of explained variance greater than 20% or compare the first and second eigenvalues of 5 or (Retnawati et al., 2017; Wells & Purwono, 2009). If the results of this assumption test show more than one dominant component, the analysis is carried out using multidimensional IRT. *Second*, testing the fit of the model. The model fit test was carried out to see whether the items followed the specified model. This analysis was assisted by using the PARSCALE program from SSi. The suitability of the model can be known from the probability value (significance, sig), provided that if the value of sig < 0.5, then the item is said not to fit the model (Retnawati, 2014). After finding the best model, the analysis continued to determine the parameter values and item characteristics based on the appropriate model. The grain parameters include the level of difficulty and discriminating power. For example, an item is good if it has a difficulty level between -2 and +2, and the discrepancy lies between 0 and 2 (Hambleton & Swaminathan, 1985). Regarding the IRT approach analysis, there are several polytomous scoring models, including the GRM, PCM, and GPCM.

Result And Discussion

Result

Before further analysis, the first test is the dimensionality test. Each item only assesses one ability when it is unidimensional (Retnawati, 2014). In contrast, multidimensional refers to objects that sometimes measure in more than one dimension. The SPSS factor analysis used in this study was used to validate the dimensional test. The KMO-MSA and Barlett tests were used as feasibility tests in the factor analysis process. The Barlett Test establishes the homogeneity of the data, whereas the KMO-MSA test seeks to determine the suitability of the sample. Factor analysis can be continued if the Kaiser Meyer Olkin (KMO)-MSA score is more than 0.5 and the Barlett s significant test is less than 0.05 (Hair et al., 2009). The KMO-SMA and Barlett scores are calculated using student answer data from the HOTS abilities instrument, as shown in Table 1. Table 1 shows that the sample used has met the sample adequacy requirements (KMO-MSA > 0.5), and the data are data homogeneous (Barlett test <0.05), so that factor analysis can be performed. Table 2 shows the data processing findings for factor analysis using SPSS in the eigenvalues component.

Volume: 8
Nomor : 4
Bulan : November
Tahun : 2022

E-ISSN: 2656-940X
P-ISSN: 2442-367X
URL: jurnal.ideaspublishing.co.id

Table 1
Result of KMO and Bartlett s Test

KMO and Bartlett s Test	
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	0.913
Approx. Chi-Square	1678.759
df	120
Sig.	0.000

Table 2
Eigenvalues

Component	Initial Eigenvalues		
	Total	% of Varians	Cumulative %
1	6.226	38.913	38.913
2	1.807	11.297	50.209
3	1.034	6.464	56.673
4	0.908	5.672	62.345
…			

Table 2 shows that eigenvalues with more than one indicate one factor; the HOTS test instrument has three factors based on this Eigenvalue. These three variables account for 38.913 of the variance. These eigenvalues are then shown in Figure1 as a scree plot.

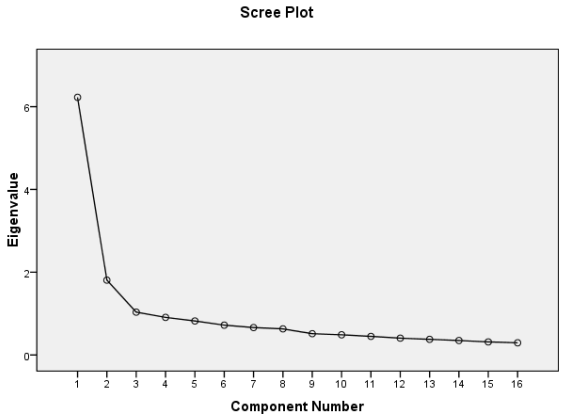

Figure 1. Scree Plot Factor Analysis

Model Fit Test

Determination of the theoretical model is based on the suitability of the instrument s character. Judging from the character of the instrument developed, GPCM is very suitable for polytomous instruments with instrument characteristics that consider each step s difficulty level to estimate participants ability (Retnawati, 2011). But to support the analysis, the model s fit must be tested. The probability value (significance, sig) can determine the model s suitability. If the value of sig < 0.5, the item is unsuitable or does not fit (Retnawati, 2014). The model containing the fittest items was selected for data analysis of the several models (GRM, PCM, GPCM2PL, GPCM3PL). The probability value (significance, sig) is obtained from the PARSCALE output. Based on the results (see Apendix) of the analysis it was found that the most suitable model is GPCM2PL.

Item Parameters

The results of the analysis of parameter estimates for HOTS instrument using the GPCM2PL IRT model can be seen in the PARSCALE phase 2 program. The results of the analysis of the estimation of the level of difficulty and different power parameters and with the GPCM2PL model for HOTS Instrument Dynamics tests are presented in Table 3.

E-ISSN: 2656-940X
P-ISSN: 2442-367X
URL: jurnal.ideaspublishing.co.id
Volume: 8
Nomor : 4
Bulan : November
Tahun : 2022

Table 3
Recapitulation of item parameters of HOTS Instrument Dynamics test

Item	Power Difference	Criteria	Difficulty index	Criteria
1	0.395	Good	-0.462	moderate
2	0.947	Good	-0.294	moderate
3	0.900	Good	-0.268	moderate
4	0.928	Good	-0.340	moderate
5	1.117	Good	-0.349	moderate
6	0.805	Good	-0.397	moderate
7	0.287	Good	-0.629	moderate
8	0.392	Good	-0.567	moderate
9	0.540	Good	-0.448	moderate
10	1.136	Good	-0.239	moderate
11	0.756	Good	-0.359	moderate
12	1.861	Good	-0.184	moderate
13	1.531	Good	-0.259	moderate
14	0.486	Good	-0.744	moderate
15	0.638	Good	-0.720	moderate
16	0.467	Good	-0.705	moderate

Ability Estimation

The ability of students to measure using the HOTS Instrument Dynamics test is shown by the amount of ability in the output of the analysis based on the GPCM2PL IRT model. The results of the analysis of students abilities on HOTS are presented in Table 4.

Table 4
Description of students abilities

Description	Value
Mean	0.000
Std.Dev	1. 002
Max	2.147
Min	-2.786

The distribution of students abilities on the two tests is presented in a histogram, as shown in Figure 2. Based on Figure 2, it can be concluded that, in general, the distribution of test takers abilities is close to the normal curve.

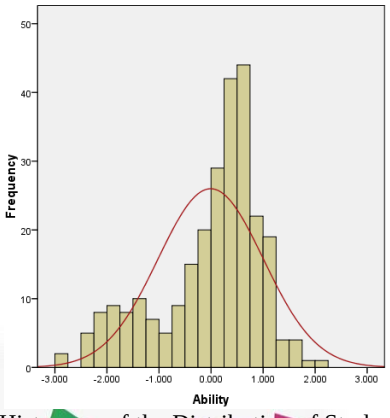

Figure 2. Histogram of the Distribution of Students Abilities

Volume: 8
Nomor : 4
Bulan : November
Tahun : 2022
E-ISSN: 2656-940X
P-ISSN: 2442-367X
URL: jurnal.ideaspublishing.co.id

Information on item characteristics and estimation of students abilities provides other information related to the test information function. If the test items have a high information function, the test information function will be high. The information function of the two test devices can be presented in Figure 3.

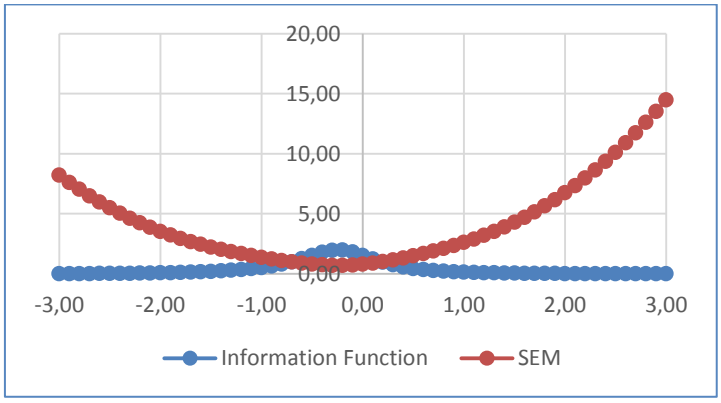

Figure 3. Information Function Tests and Standard Error of Measurement (SEM)

Discussion

The first stage in data analysis is the dimensionality test. Factor analysis is used to evaluate its dimensionality, starting with the KMO test to guarantee the sample s sufficiency. It uses the KMO test. The sample adequacy test determines whether or not the sample obtained meets the sample adequacy criteria (KMO-MSA > 0.5), and the data are data homogeneous (Barlett test <0.05), so that factor analysis can be performed. The unidimensionality test using factor analysis and scree plot found that, from these two scree plots, the Eigenvalue appears to decline sharply between factors 1 and 2; the Eigenvalue then begins to tilt at factor 3 so that the scree plot almost forms a right-angled angle. The HOTS instrument test tool measures at least two dominant factors. Although it appears to measure two factors, if you look at the percentage of variance explained by the first factor for the two test sets (38,913), the value is greater than 20%. In line with this, comparing the first with the second eigenvalues shows five times. Both of these conditions meet the requirements to be said to be (Heri Retnawati, 2017; Wells & Purwono, 2009). Based on this test, it can be concluded that instrument sets only contain a single dimension or are unidimensional. Local independence is another test. This premise of local independence will be met if the participant s response to one item does not affect the participant s response to the other items (Retnawati, 2014). According to De Mars (2010) (Salkind, 2013), the unidimensional assumption can also be used to discover local independence (Retnawati, 2016). The local independence assumption is also met if the unidimensional assumption is met. Because the unidimensional assumption was met in this investigation, the local independence test was also met.

Based on analysis for model Fit Test, the results show that the fittest or provide information on each item are GPCM2PL. This result is 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 (Hidayati & 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 scored in a tiered category. Still, the difficulty index in each step is not ordered; a step can be more difficult than the next step. Therefore, Istiyono asserted that using GPCM to analyze multiple-choice tests is a fair alternative assessment model in learning (Istiyono et al., 2020).

Further analysis for item parameters shows that the recapitulation of the results of the analysis shows that for items of the HOTS test, the different power parameter value s 100% is included in the good category, namely the interval 0.00 to 2.00 (Hambleton & Swaminathan, 1985) and the difficulty level analysis also shows the percentage of 100% of items included in the medium category because the b value of all items is in the interval (-2) to (2) (Hambleton & Swaminathan, 1985). It shows that the items that make up the test are worthy of being used as a good instrument and can accurately measure students abilities. The feasibility of the instrument in measuring student abilities can be viewed from the information function

Based on the result, the HOTS test provides the highest information for students with abilities around -0.3. It is also characterized by the smallest standard error in the range of capabilities. In the interval -0.7 to + 0.3, the information function s value is greater than the standard error measurement (SEM) so that the measurement accuracy is considered good (Retnawati, 2014), and the smaller the SE, the greater the reliability of the test (Salkind, 2013). Based on this information, the HOTS instrument accurately measures students ability (θ) between this interval of -0.7 to +0.3.

Conclusions

Assessment using test instruments needs to consider the characteristics of the test instruments used. This research attempts to analyze the characteristics of the HOTS test instrument items using the IRT model of polytomous scoring that is fit. Based on the fit test, it was found that the model that is most suitable for the data is GPCM2PL. The item parameter analysis shows that the value of the different power parameters for all items with a percentage of 100% is included in the good category, namely the interval of 0.00 to 2.00. The difficulty level analysis shows the percentage of 100% of items included in the medium category because the b value of all items is in the interval (-2) to (2).

References

Ab Kadir, M. A. (2017). What Teacher Knowledge Matters in Effectively Developing Critical Thinkers in the 21 st Century Curriculum? *Thinking Skills and Creativity*, *23*, 79–90. https://doi.org/10.1016/j.tsc.2016.10.011

Adeyemo, S. a. (2010). Students Ability Level and Their Competence in Problem-Solving Task in Physics. *International Journal of Educational Research and Technology*, *1*(December), 35–47. http://www.soeagra.com/ijert/vol2/7.pdf

Ahrari, S., Samah, B. A., Hassan, M. S. H. Bin, Wahat, N. W. A., & Zaremohzzabieh, Z. (2016). Deepening Critical Thinking Skills Through Civic Engagement in Malaysian Higher Education. *Thinking Skills and Creativity*, *22*, 121–128. https://doi.org/10.1016/j.tsc.2016.09.009

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy of Learning, Teaching, and Assessing: a Evisrion of Bloom s Taxonomy of Educational Objectives*. Longman.

Asysyifa, D. S., . J., Wilujeng, I., & Kuswanto, H. (2019). Analysis of Students Critical Thinking Skills Using Partial Credit Models (PCM) in Physics Learning. *International Journal of Educational Research Review*, *4*(2), 245–253. https://doi.org/10.24331/ijere.518068

Bano, V. O., Marambaawang, D. N., & Njoeroemana, Y. (2022). Analisis Kriteria Butir Soal Ujian Sekolah Mata Pelajaran IPA di SMP Negeri 1 Waingapu. *Ideas: Jurnal Pendidikan, Sosial, dan Budaya*, *8*(1), 145. https://doi.org/10.32884/ideas.v8i1.660

Behar-Horenstein, L. S., & Niu, L. (2011). Teaching Critical Thinking Skills In Higher Education: A Review Of The Literature. *Journal of College Teaching & Learning (TLC)*, *8*(2). https://doi.org/10.19030/tlc.v8i2.3554

Bloom, B. S. (1972). Taxonomy of Educational Objectives: The Classification of Educational Goals, Parts 1-2. In *Handbook I: Cognitive Domain.* Ann.

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

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

Cotton, K. (1997). Developing Employability Skills. In *School Improvement Research Series* (Issue 1987). http://www.nwrel.org/scpd/sirs/8/0015.html

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.

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An Integrated Critical Thinking Framework for the 21st Century. In *Thinking Skills and Creativity* (Vol. 12, pp. 43–52). https://doi.org/10.1016/j.tsc.2013.12.004

Guo, J., & Woulfin, S. (2016). Twenty-First Century Creativity: An Investigation of How the Partnership for 21st Century Instructional Framework Reflects the Principles of Creativity. *Roeper Review*, *38*(3), 153–161. https://doi.org/10.1080/02783193.2016.1183741

Hadi, S., Retnawati, H., Munadi, S., Apino, E., & Wulandari, N. F. (2018). The Difficulties of High School Students in Solving Higher-Order Thinking Skills Problems. *Problems of Education in the 21st Century*, *76*(4), 520.

E-ISSN: 2656-940X
P-ISSN: 2442-367X
URL: jurnal.ideaspublishing.co.id

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* (Vol. 2). Sage.

Hasan, M., Maulidyanti, H., Tahir, M. I. T., & Arisah, N. (2022). Analisis Keterampilan Berpikir Kritis Peserta Didik Melalui Kegiatan Literasi. *Ideas : Jurnal Pendidikan, Sosial, dan Budaya*, *8*(2), 477–486.

Hasudungan, A. N., & Kurniawan, Y. (2018). Meningkatkan Kesadaran Generasi Emas Indonesia dalam Menghadapi Era Revolusi Industri 4.0 Melalui Inovasi Digital Platform www.indonesia2045.org. *Prosiding Seminar Nasional Multidisiplin*, *1*, 51–58. https://ejournal.unwaha.ac.id/index.php/snami/article/view/263

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

Hidayati, K., & Retnawati, H. (2011). *Identifying Item Bias Of Test Using The Probability Difference Indices On Polytomous Data Using Generalized Partial Credit Model*.

Istiyono, E. (2017). The Analysis of Senior High School Students Physics HOTS in Bantul District Measured Using PhysReMChoTHOTS. *AIP Conference Proceedings*, *1868*(1), 70008. https://doi.org/10.1063/1.4995184

Istiyono, E. (2019). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika SMA. *Jurnal Inovasi dan Pembelajaran Fisika*, *6*(1), 70–81. https://doi.org/10.36706/jipf.v6i1.7817

Istiyono, E., 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

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.

King, F. J., Goodson, L., & Rohani, F. (2010). *Higher Order Thinking Skills: Definition, Teaching Strategies, Assessment*. http://goo.gl/su233T.

Kizlik, B. (2012). Measurement, Assessment, and Evaluation in Education. In *FGS, UiTM* (pp. 1–43). http://drjj.uitm.edu.my

Krulik, S., & Rudnick, J. A. (1999). Innovative Tasks to Improve Critical and Creative Thinking Skills. In D. L. V Stiff & F. R. Curcio (Eds.), *Developing Mathematical Reasong in Grades K-12* (pp. 138–145). NCTM.

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

Liu, S., Yang, X., Zhang, H., Wang, Y., Yoneda, T., & Li, Z. (2018). Study on Teaching Methods for Developing Higher Order Thinking Skills for College Students in Flipping Classroom. *Proceedings - 6th International Conference of Educational Innovation Through Technology, EITT 2017*, *2018-March*, 254–257. https://doi.org/10.1109/EITT.2017.69

Marlina, L., Liliasari, Tjasyono, B., & Hendayana, S. (2018). Improving the Critical Thinking Skills of Junior High School Students on Earth and Space Science (ESS) Materials. *Journal of Physics: Conference Series*, *1013*(1), 12063. https://doi.org/10.1088/1742-6596/1013/1/012063

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

Prakash, R., & Litoriya, R. (2022). Pedagogical Transformation of Bloom Taxonomy s LOTs into HOTs: An Investigation in Context with IT Education. *Wireless Personal Communications*, *122*(1), 725–736.

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.

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*(48–60).

Republik Indonesia. (2012). *Peraturan Presiden RI (Perpres) No. 8 Tahun 2012 tentang Kerangka Kualifikasi Nasional Indonesia*. Indonesia.

Retnawati, H. (2011). Mengestimasi Kemampuan Peserta Tes Uraian Matematika dengan Pendekatan Teori Respons Butir dengan Penskoran Politomus dengan Generalized Partial Credit Model. In *Prosiding Semnas Penelitian Pendidikan dan Penerapan MIPA. UNY* (pp. 53–62).

Retnawati, H. (2014). Teori Respons Butir dan Penerapannya: Untuk Peneliti, Praktisi Pengukuran dan Pengujian, Mahasiswa Pascasarjana. In *Yogyakarta: Nuha Medika*. Parama. Retnawati, H.

Retnawati, H. (2016). Analisis Kuantitatif Instrumen Penelitian (Panduan Peneliti, Mahasiswa, dan Psikometrian). In *Parama Publishing*. Parama Publishing.

Retnawati, H., Hadi, S., Nugraha, A. C., Arlinwibowo, J., Sulistyaningsih, E., Djidu, H., Apino, E., & Iryanti, H. D. (2017). Implementing the Computer-Based National Examination in Indonesian Schools: The challenges and strategies. *Problems of Education in the 21st Century*, *75*(6), 612–633. https://doi.org/10.33225/pec/17.75.612

Robinson, J. P. (2000). A Fact Sheet: What Are Employability Skills? In *Alabama Cooperative Extension System* (Vol. 1, Issue 3). http://proquest.umi.com/pqdweb.

Rosmiati, R., & Satriawan, M. (2019). The Ocean Climate Phenomenon: The Challenges of Earth Physics Lectures in Indonesia. *Journal of Physics: Conference Series*, *1157*(3), 32038. https://doi.org/10.1088/1742-6596/1157/3/032038

Salkind, N. (2013). Item Response Theory. In *Encyclopedia of Measurement and Statistics*. Oxford University Press. https://doi.org/10.4135/9781412952644.n230

Saprudin, S., Liliasari, S., Prihatmanto, A. S., & Setiawan, A. (2019). Pre-service Physics Teachers Thinking Styles And Its Relationship with Critical Thinking Skills On Learning Interference and Diffraction. *Journal of Physics: Conference Series*, *1157*(3), 32029. https://doi.org/10.1088/1742-6596/1157/3/032029

Satriawan, M., Liliasari, W., & Abdullah, A. G. (2020). Analysing of Pre-Service Physics Teachers Critical Thinking Skills Profile in Ocean Wave Energy Topic. *Journal of Physics: Conference Series*, *1521*, 22041.

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

Wells, C. S., & Purwono, U. (2009). *Assesing the Fit of IRT Models to Item Response Data*. Makalah Pelatihan Psikometri Kerjasama Pascasarjana UNY dengan USAID.

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