Characteristics of Two-Tier Multiple Choice (TTMC) High Order Thinking Skill (HOTS) Instruments and the Estimation of HOTS Abilities of Vocational Education Students Using the Item Response Theory Approach

Duden Saepuzaman[1], Edi Istiyono[2], Haryanto[2], Heri Retnawati[2]

[1]Universitas Pendidikan Indonesia, Bandung, Indonesia
[2]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[1]dsaepuzaman@upi.edu,
[2]edi_istiyono@uny.ac.id, haryanto@uny.ac.id, heri_retnawati@uny.ac.id

***Abstract*- This research aims to design and build an isomorphous binary phase diagram animation media using Adobe Flash Abstract—This study aims to determine the characteristics of the HOTS instrument item parameters in the form of two-tier multiple-choice (TTMC) and estimate the HOTS ability of vocational high school students using the Item Response Theory (IRT) approach. The research method used is quantitative. The research subjects were 264 vocational high school students using class XI in West Java and Banten provinces. The test instrument used was Newton's Law concept and Two-Tier Multiple Choice (TTMC). Data analysis includes a unidimensional test, the fitness of the polytomous IRT model, the determination of item characteristics, and the estimation of students' HOTS ability. The results showed that the data were unidimensional, and the fitness IRT polytomous scoring model was GPCM2PL. The discrimination parameter (a) of all items with a percentage of 100% is included in the good criteria. Analysis of the level of difficulty (b) also shows that 100% of items are included in the moderate criteria. This result was obtained because the instrument used was developed through developing an appropriate instrument. Thus, its validity and reliability were good. It means that the instrument used is feasible to be tested on vocational high school students. Other studies show that the mean of students' HOTS abilities is in the medium category.**

***Keywords*- TTMC Instruments, HOTS, vocational high schools, Item Response Theory (IRT).**

Duden Saepuzaman
Universitas Pendidikan Indonesia, Indonesia
dsaepuzaman@upi.edu

I. INTRODUCTION

The vocational school curriculum is designed in such a way as to prepare a professional workforce and prepare students to anticipate future needs and challenges that are aligned with the development of the needs of the business /industry, the development of the world of work, and the development of science and technology (Rosina et al., 2021; Maryanti & Nandiyanto, 2021).

Physics is one of the branches of science that underlies the development of science and technology. Physics competence which is expected to meet these demands is to become the foundation for vocational competencies. To keep up with the development of science and technology, it is hoped that vocational students will be equipped with hard skills and need to have higher-order thinking skills (Luthvitasari & Linuwih, 2012). Many reports relating to physics education have been well-documented (Susilowati et al., 2023; Lestari et al., 2024; Abosede et al., 2024; Azizah et al., 2024; Ibrahim, 2023; Al Husaeni, 2022).

Some experts associate high order thinking skills (HOTS) with a variety of thinking skills that can be performed by individuals. According to experts, thinking skills that can be categorized as HOTS include critical thinking and creative thinking skills (Conklin, 2012: 14; Presseisen, 1988; Krulik & Rudnick, 1999: 138; King, Goodson, & Rohani, 2010: 1), problem-solving (Presseisen, 1988; Brookhart: 2010: 3), logical, reflective, and metacognitive thinking (King, Goodson, & Rohani, 2010: 1), and decision making (Presseisen, 1985: 46). These skills are not foreign terms in the learning process; they have even become a target and part of the learning objectives in each

subject (Jailani et al., 2018). This research will focus on HOTS, which includes indicators of critical thinking skills and creative thinking skills. It is because it is in line with the core competencies of vocational school.

One of the studies that are very close to HOTS is Physics (Adeyemo, 2010). Apart from being part of the competence of vocational students, this is also because there is a significant relationship between HOTS and student performance (Ramos et al., 2013). HOTS will help students understand concepts more easily, be sensitive to problems that occur, understand and solve problems that occur around them, and apply these concepts in different situations (Marlina, L., Tjasyono, B., & Hendayana, S., 2018; Rosmiati, R., & Satriawan, M. 2019). Besides being very close to the performance, HOTS is also part of 21st-century skills (Guo & Woulfin, 2016; Dwyer CP, Hogan MJ and Stewart I 2014; Saprudin S, Liliasari S, Prihatmanto A S and Setiawan A 2019; Ahrari, et al. 2016; Ab Kadir, MA, 2017).

To determine the extent to which the HOTS achievement of vocational school students needs to be assessed. Assessment is the process of gathering information related to learning objectives or outcomes (Kizlik, 2012: 7). Even though HOTS is one of the learning objectives, the facts in the field show that HOTS assessments in schools are still not fully carried out. The assessments carried out in schools are still predominantly limited to the memory aspect. The cause of this problem is that they are still limited in working on HOTS-based questions, and teachers are still lacking in developing HOTS instruments (Budiman & Jailani, 2014). Even if there is a HOTS instrument, the instrument used has not been tested for its feasibility and characteristics, such as distinguishing power and difficulty level.

The test instrument commonly used is multiple choice and focuses on aspects of knowledge. Multiple choice questions are most commonly used because it is considered very easy to apply and easy to analyze. This form is often criticized for being only able to assess shallow memorization or simple facts because it does not allow test takers to explain or justify their answers (Nichols & Sugrue, 1999; Songer, Kelcey & Gotwals, 2009), although in some cases this weakness can reduce (Xiao, et al. 2018; Hestenes, Wells & Swackhamer 1992).

The development of the type of reasoned multiple-choice questions (Reasoning Multiple Choice) is seen as measuring high-level abilities or skills (Xiao et al. 2018; Liu et al., 2011). The same aspect was expressed by Cullinane & Liston (2011) that the inclusion of reasons at the second level of the two-tier choice question form can be used to improve higher-order thinking skills and see the ability of test-takers to give reasons. Thus, in choosing the answer, the test taker must think about the reasons that are following the choice of the answer, directly the thought process of determining the right reason can train the test taker's higher-order thinking skills. In addition, it can also be seen that the lack of quality assessment is due to the choice of a multiple-choice test model that is commonly used to measure low-level thinking skills (Istiyono, Mardapi, Suparno, 2014). According to Brookhart (2010: 33), multiple-choice tests must be modified. Thus, they can be used to measure higher-order thinking skills. One of the efforts made is to make a two-tier instrument called a two-tier multiple-choice (TTMC) (Istiyono et al., 2014; 2020a; 2020b). Regarding the 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 try and ensure that the assessment results accurately describe the student's abilities. An assessment is called accurate if the results of the assessment contain the smallest possible error or error. To get results that accurately describe the abilities of students, the quality of the test instruments must be valid, reliable, and have good item parameters. For this purpose, two types of approaches can be used to estimate item parameters, namely classical test theory and item response theory. The classical test theory is seen to have weaknesses. The most notable weakness of classical test theory is that the characteristics of the examinee and the characteristics of the test are inseparable, each of which can only be interpreted in another context (Hambleton, Swaminathan, & Rogers, 1991). That is, the ability of the examinees is determined only by the test. When the test is difficult, the examinee will appear to have low ability, and when the test is easy, the examinee will appear to have a higher ability. In other words, the parameter of the item depends on the subject/test taker and vice versa. The characteristics of the items will change when the examinee changes 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 are very dependent on the test taker subject.

Item response theory (IRT) is a solution to overcoming weaknesses in classical test theory because item response theory has the concept of releasing the relationship between items and samples or test taker subjects. The characteristics/abilities of the examinees will remain the same even though they are working on items with different characteristics and vice versa; the characteristics of the items will remain the same even though test takers perform them with different abilities. In addition, the item response theory is based on items, not on test kits. According to Hambleton et al. (1991), item response theory rests on two postulates: (a) a test taker's performance on test items can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and (b) the relationship between the test taker's performance and the item can be explained by a function that increases monotonically called the item characteristic curve (ICC) function which is presented in Figure 1. This function explains that when the ability increases, the respondent's probability of answering correctly for an item also increases. Figure 1 shows the group of test-takers with a higher ability

will have a greater probability of answering correctly than the group of test-takers with low ability. The function of item response theory can be applied when the model is compatible with the test data (Hambleton et al., 1991). Stone & Zhang (2003) stated that grain parameter estimation could be disturbed when the model does not match the data. 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).

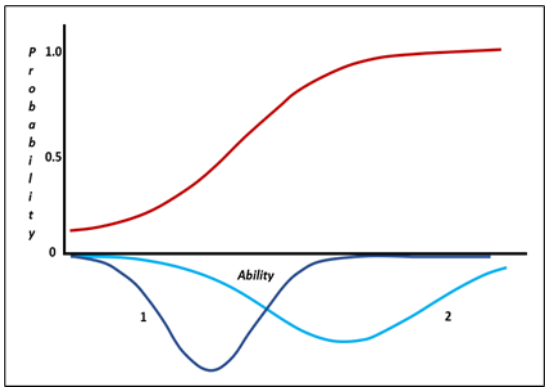

Fig. 1. ICC curves and ability distributions in two groups of test-takers (adapted from Hambleton et al., 1991)

In the Graded Response Model (GRM), participants' responses to item j with the GRM model are categorized into m + 1 ordered category scores, k = 0,1,2,…, m where m is the number of steps completing item j correctly, and the difficulty index. Each step is also sorted. The relationship between item parameters and participant abilities in GRM for homogeneous cases (same in each step) can be stated by Muraki & Bock (in Retnawati, 2014) as shown in equations (1) dan (2).

$$P_{jk}(\theta) = P_{jk}^{*}(\theta) - P_{j\,k+1}^{*}(\theta) \quad (1)$$

$$P_{jk}(\theta) = \frac{exp\,[Da_j(\theta - b_{jk})]}{1 + exp\,[Da_j(\theta - b_{jk})]} \quad (2)$$

where
a_j : Discriminationindex item j,
θ : the ability of the participant,
b_{jk} : category k difficulty index item j,
$P_{jk}(\theta)$: the probability of a capable participant θ who gets the k category score in point j,
$P_{jk}^{*}(\theta)$: the probability of a capable participant θ obtaining a category k score or more on item j,
D : scale factor,

Partial Credit Model (PCM) is a polytomous scoring model that extends the Rasch model in dichotomous data. The assumption on PCM is that each item has the same distinctive power. PCM has similarities with the Graded Response Model (GRM) in the items that are scored in the tiered category, but the difficulty index in each step does not need to be sorted, one step can be more difficult than the next step. According to Muraki & Bock (1997: 16), the general form of PCM is equation (3).

$$P_{jk}(\theta) = \frac{expexp\,[Da_j(\theta - b_{jk})]}{1 + expexp\,[Da_j(\theta - b_{jk})]} \quad (3)$$

$k = 0,1,2, \ldots . m$
where
$P_{jk}(\theta)$: the probability of capable participant θ who gets the k category score in point j
θ : participant's ability
$m + 1$: number of categories of item j
b_{jk} : category k difficulty index item j

The category score on the PCM indicates the number of steps to complete the item correctly. A higher category score indicates greater ability than a lower category score. Suppose an item has two categories in PCM. In that case, the equation will become the Rasch model equation, such as the equation stated by Hambleton, and Swaminathan (1985), and also reinforced by Hambleton, Swaminathan, and Roger (1991). As a result, PCM can be applied to polytomous and dichotomous items (Retnawati, 2014).

According to Muraki (1997), GPCM is a general form of PCM, which is expressed in mathematical form, which is called the item category response function expressed in equations (4) and (5) (Retnawati, 2014).

$$P_{jk}(\theta) = \frac{\exp \sum_{v=0}^{k} Z_{jk}(\theta)}{\sum_{e=0}^{m_j} \exp(\sum_{v=0}^{e} Z_{jk}(\theta))} \quad (4)$$

and

$$Z_{jk}(\theta) = Da_j(\theta - b_{jk}) = Da_j(\theta - b_j + d_k) \quad (5)$$

where
$P_{jk}(\theta)$: the probability that the participant with the ability gets the k category score in item j,
θ: the ability of the participant ,
a_j : index of difference in item j,
b_{jk}: category k difficulty index in item j,
b_j : location difficulty index item j abbreviated as an item location parameter,
d_k : k category parameter
m_j : number of categories of item j,
D : scale factor (D = 1.7) ,

The parameter b_{jk} by Masters (Muraki & Bock, 1997: 17) is named with the grain stage parameter. This parameter is the intersection point between the Pjk(θ) and Pjk-1(θ).) curves. The two curves only intersect at one point on the θ scale (van der Linden & Hambleton, 1997).

GPCM is formulated based on the assumption that each probability of choosing the k-th category exceeds the (k-1) category is constructed by a dichotomous model. Pjk is the specific probability of choosing the k-th category from the mj + 1 category. The relationship between the probability of getting correct for each ability θ is presented in a Categorical

Response Function (CRF) graph (du Toit, 2003; Retnawati, 2014).

Based on the explanation above, this study focuses on analyzing the characteristics of the item parameters and estimating the HOTS ability of vocational students using the Item Response Theory (IRT) approach.

II. METHOD

This study used a quantitative approach. The research subjects were 264 students of class XI in vocational schools in West Java and Banten provinces, Indonesia. The instrument used was a HOTS instrument in Newton's Law of Motion and Force, amounting to 30 items in Two-Tier Multiple Choice (TTMC) with four categories. The instrument used has been tested for validity and reliability.

For example, the instrument item (number 3) used is presented in Figure 2.

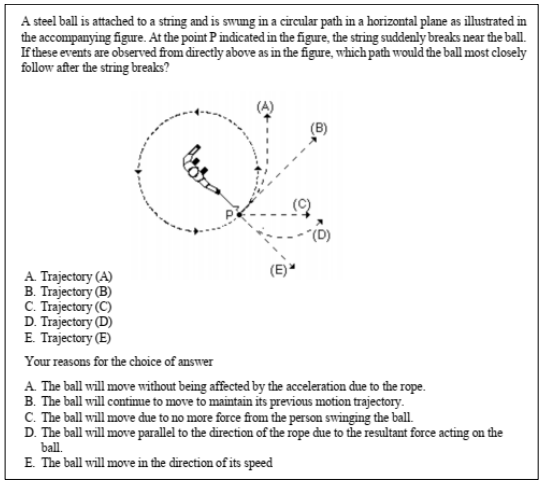

Fig. 2. Examples of Two-Tier Multiple Choice (TTMC) Questions

The scoring criteria used are presented in Table 1 (Istiyono et al, 2014; 2020a; 2020b). With the TTMC instrument and scoring form as in Table 1, the collected data is polytomous.

TABLE I
TEST SCORING CRITERIA

Criteria	Case
4	The answer to the question and the reason is correct
3	The answer to the question is wrong but the reason is right
2	The answer to the question is correct but the reason is wrong
1	Wrong answers to questions and reasons

The instrument developed refers to the HOTS indicator, which follows the competence of vocational students. The HOTS indicator developed refers to indicators of critical thinking skills (Ennis, 1985; Facione, 2011; Facione & Gittens, 2015; Abrami, et.al. 2008) and creative thinking skills (Torrance & Shaughnessy, 1988; Haefele (1962); Tahar, Tej & Sirkova, 2015; Ferrandiz et.al., 2017) which is presented in Table 2.

TABLE II
HOTS INDICATORS

Indicator	Item
Identify/formulate questions based on events in everyday life	1, 16
Analyze statements and determine the similarities or differences of a given event	2, 17
Test/examine sections that can be considered to be trustworthy (or untrustworthy) based on the text of the argument, advertisement, or experiment and its interpretation, and provide reasons why	3, 18
Reveal reasons based on observations of an event	4, 19
Interpret statements and clarify data	5, 20
Generalize (finding patterns) based on existing data trends	6, 21
Solve problems by Using definitions	7, 22
Formulate alternatives for solutions	8, 29
Answer with many answers or facts	9, 24
Look at the faults and weaknesses of an object or situation	10, 25
Provide interpretation of a picture, story, or problem	11, 26
Think of a way or point of view to solve the problem	13, 27
Classify things according to different divisions (categories)	12, 28
Develop or enrich the ideas of others	14, 30
Try/test new things by experimenting	15, 23

The data analysis carried out included (i) unidimensional testing using SPSS, (ii) determining the fitness of the Politomus IRT model using the PARSCALE of SSi (iii) Analysis of the characteristics of the instrument items based on good item criteria and (iv) estimating the HOTS ability of vocational school students based on the IRT approach. Detailed information regarding the use of statistical analysis is explained elsewhere (Fiandini et al., 2024; Afifah et al., 2022).

III. 3. RESULTS AND DISCUSSION

3.1. Unidimensional Assumption

Before the fit test stage of a fit polytomous model, the first thing to do is the assumption test. This assumption test includes tests of local unidimensional and independence. Unidimensional means that each item measures only one ability (Retnawati, 2014). While multidimensional means that some or all items measure more than one dimension. The dimensional test in this study was proven through factor analysis using SPSS. Factor analysis was done by first doing 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 Barlett's significant test <0.05 (Hair, JF, Black, WC, Babin, BJ, Anderson, RE, & Tatham, RL, 2009). Based on the response data of students on the HOTS instrument, the

KMO-SMA and Barlett values were obtained as presented in Table 3.

TABLE III
KMO AND BARTLETT'S TEST

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**0.904**
Bartlett's Test of Sphericity	Approx. Chi-Square	3509.819
	df	435
	Sig.	0.000

Table 3 shows that the sample used has met the adequacy requirements of the sample (KMO-MSA> 0.5) and the data is homogeneous (Barlett's test <0.05). Thus, factor analysis can be carried out. The results of data processing for factor analysis through SPSS can be seen in the eigenvalues section of TABLE 4. Based on TABLE 4, the eigenvalues with more than one value indicate one factor. Based on these eigenvalues, the HOTS instrument has seven factors. These seven factors can explain the 59.763% variance. These eigenvalues can then be presented in the scree plot in Figure 3.

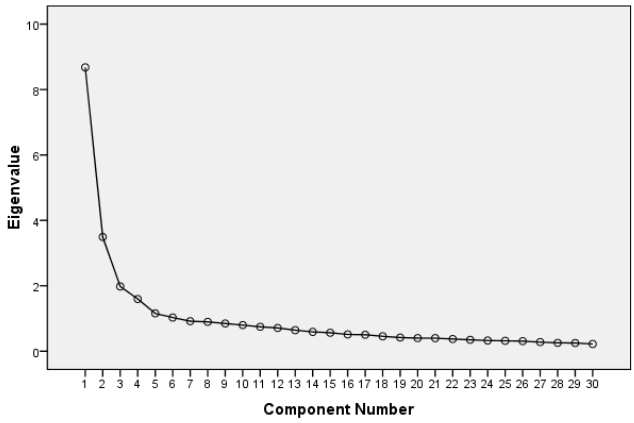

Fig. 3. Scree Plot Factor Analysis

Figure 3 shows a very sharp decrease between factor 1 and factor 2. The Eigenvalue then begins to skew at a factor of 3. Thus, the scree plot almost forms a right angle. It shows that there is only 1 dominant factor in the two test sets. This follows the statement of Hambleton, Swaminthan & Rogers (1991: 56) that the unidimensional assumption can be considered fulfilled if the test contains a dominant component that measures students' ability. In addition, Wells and Purwono (Retnawati, 2016: 144) confirm the amount of the explained variance presentation. If the value is greater than 20%, the measured device contains a single dimension or is unidimensional.

Another test is local independence. This assumption of local independence will be fulfilled if the participant's answer to one item does not affect another item's answer (Retnawati, 2014). According to De Mars (2010), local independence can also be detected by proving unidimensional assumptions. This means that if the unidimensional assumptions are met, the local independence assumption is also fulfilled ((Retnawati, 2014). This is because if the data that is owned is unidimensional, the response given by the test taker to an item is independent or does not affect the test taker's answer to the item. This means that the test taker's ability is independent of the HOTS instrument items. In this study, the unidimensional assumptions have been fulfilled. Thus, the local independence test has also been fulfilled.

3.2. Fitness of The Politomus IRT Model

Model fit can be seen from the probability value (significance, sig). If the sig value < α, then the items are said to be unsuitable (Retnawati, 2014, page 25). The model containing the most fit 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. The suitability of each HOTS instrument item with the GRM, PCM, GPCM2PL, and GPCM3PL models is presented in the appendix. The result shows the IRT model that is most fit or provides information on each item is GPCM2PL.. Thus, it can be said that the IRT model with suitable polyatomic data is used to analyze the items of the HOTS test, namely GPCM2PL. This result is in line with Si (2004: 173), who 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 these items are stored in a tiered category, but the difficulty index in each step is not sorted, a one-step can be more difficult than the next. Istiyono (2016: 29) emphasizes that the use of GPCM to analyze multiple-choice tests is a fair alternative assessment model in learning.

3.3. Characteristics of The Instrument Items

The results of the analysis of the HOTS instrument item parameter estimation using the IRT GPCM2PL model can be seen in the PARSCALE phase 2 program. The results of the analysis of the parameter estimation of Discrimination and level of difficulty with the GPCM2PL model for the HOTS test are presented in TABLE 4.

Table 4 shows that the value of the different power parameters (a) of all items with a percentage of 100% is included in the good category, namely the interval from 0.00 to 2.00. The difficulty level analysis (b) also shows that 100% of items are included in the medium category because the b value of all items is in the interval (-2) to (2). It means that the HOTS instrument has good characteristics. These characteristics are very important, given the role of the test instrument to be able to measure the ability of test-takers as accurately as possible, distinguishing test-takers whose abilities are low, medium, or high.

The relationship between the probability of getting correct for each ability θ is presented in a Categorical Response Function (CRF) graph (du Toit, 2003; Retnawati, 2014). CRF charts in four categories (scores 1,2,3, and 4) for item number 3 with a difference of 0.807 and a difficulty level of -0.386 are presented in Figure 4.

TABLE IV
RECAPITULATION OF HOTS INSTRUMENT ITEM PARAMETERS

Item	Discrimination	Criteria	Difficulty Level	Criteria
1	0.395	good	-0.462	moderate
2	0.947	good	-0.294	moderate
3	0.807	good	-0.386	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	0.357	good	-0.588	moderate
10	0.329	good	-0.707	moderate
11	1.181	good	-0.486	moderate
12	0.941	good	-0.261	moderate
13	0.487	good	-0.336	moderate
14	0.900	good	-0.268	moderate
15	0.147	good	-1.547	moderate
16	0.540	good	-0.448	moderate
17	1.136	good	-0.239	moderate
18	0.756	good	-0.359	moderate
19	1.861	good	-0.184	moderate
20	1.531	good	-0.259	moderate
21	0.486	good	-0.744	moderate
22	0.638	good	-0.720	moderate
23	0.434	good	-0.897	moderate
24	0.949	good	-0.373	moderate
25	0.819	good	-0.502	moderate
26	1.819	good	-0.519	moderate
27	1.289	good	-0.349	moderate
28	1.360	good	-0.452	moderate
29	0.467	good	-0.705	moderate
30	0.291	good	-0.107	moderate

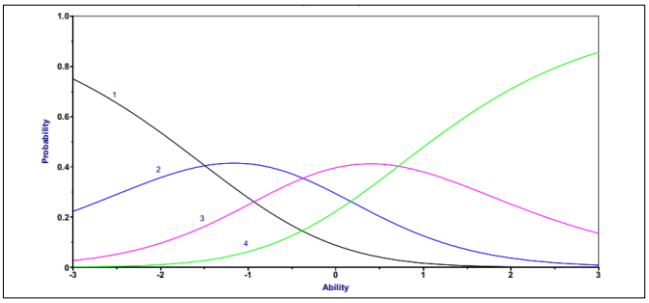

Fig. 4. CRF Graph on 4 Categories for Item Number 3

Based on Figure 4, the intersection point between the black line and the blue line (1 and 2), states the minimum opportunity and ability or is often referred to as the student's step (δ_{12}) parameter to get a score of 2. In the picture, it shows the ability score of -1.50 with a chance of 0.4. This means that students who have abilities below -1.50 have a great chance of getting a score of 1 with opportunities ranging from 0.4 to 1 (the chance of getting a value of 1 decreases with the size of the ability). Students who have the ability of -1.50 have a chance to get a score of 1 or 2 of 0.44. Students whose abilities are above -1.50 to -0.30 have a chance to get a score of 2 (δ_{23}) as well as for the Kurav intersection in the next category.

3.4. Estimating The HOTS Ability

The HOTS ability of students is shown by the ability to output the analysis results based on the GPCM2PL IRT model. The distribution of students' HOTS abilities in the range -3 to +3 (IRT model) is presented in Figure 5.

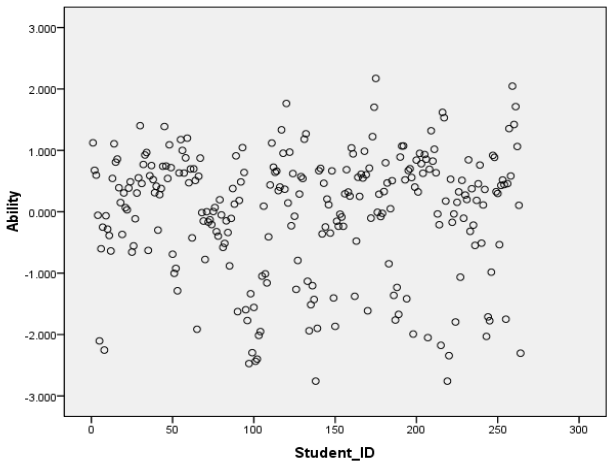

Fig. 5. Distribution of Students' HOTS Abilities

Figure 5 shows that the distribution of students' abilities tends to be slightly above average. This can be seen from the distribution of students' abilities which are in the zero ability range. This distribution is clearer in the histogram display of the distribution of students' HOTS abilities in Figure 6.

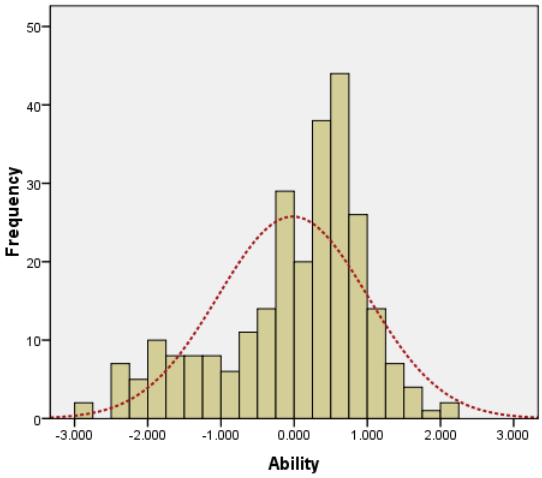

Fig 6. Histogram of Distribution Students' HOTS Abilities

Figure 6 shows the tendency of students' HOTS abilities to be at average ability. This data is also strengthened by the statistical descriptions presented in Table 8.

TABLE V
STATISTICAL DESCRIPTION OF STUDENTS' HOTS ABILITIES

Statistical Description	Magnitude
Mean	0.000
Std.Dev	0.999
Max	2.172
Min	-2.759

Table 5 shows the average HOTS ability of students is 0,000 with a standard deviation of almost the same, namely 1. It indicates that the student's abilities are classified as medium or average in general, while the standard deviation size shows that the student's abilities are quite varied. Information on student abilities is very important to make planning and appropriate learning treatment. Because of the learning that is held, for example, learning to train HOTS, it would be more appropriate if the teacher knew the HOTS abilities of students. This is in line with Hermawan (2014) that learning outcomes will be maximized if the learning is carried out following the characteristics of students

CONCLUSION

Vocational high school students' competencies are related to hard skills and the need to have thinking skills as part of the demands of 21st-century competencies, one of which is HOTS. To determine HOTS achievement, an assessment is required. Assessment is usually done through a test instrument. Through this research, the student's HOTS instrument in the form of TTMC has been analyzed which has good characteristics. Thus, it can describe the actual abilities of the students. Through the IRT approach, the characteristics of the good and suitable items to be used to assess students' HOTS and measure students' abilities were obtained. Based on these results, the ability of vocational school students is in the medium category or have average abilities. The implication is that it takes an effort to increase the student's HOTS following the students' abilities or characteristics. Thus, in the end, it can be planned and applied in learning to improve student learning outcomes. One further research that can be done is finding the right learning model for a class where most students have moderate HOTS abilities. Thus, the learning outcomes are more optimal.

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.

Abosede, P.J., Onasanya, S.A., and Ngozi, O.C. (2024). Students self-assessment of demonstration-based flipped classroom on senior secondary school students' performance in physics. Indonesian Journal of Teaching in Science, 4(1), 27-40.

Adeyemo, S. A. (2010). Students ability level and their competence in problem solving task in physics. International Journal of Educational Research and Technology, 1(2), 35-47.

Afifah, S., Mudzakir, A., and Nandiyanto, A.B.D. (2022). How to calculate paired sample t-test using SPSS software: From step-by-step processing for users to the practical examples in the analysis of the effect of application anti-fire bamboo teaching materials on student learning outcomes. Indonesian Journal of Teaching in Science, 2(1), 81-92.

Ahrari, S., Samah, B. A., Hassan, M. S. H. B., 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.

Aizikovitsh-Udi, E., & Amit, M. (2011). Developing the skills of critical and creative thinking by probability teaching. Procedia-Social and Behavioral Sciences, 15, 1087-1091.

Akpur, U. (2020). Critical, Reflective, Creative Thinking and Their Reflections on Academic Achievement. Thinking Skills and Creativity, 37, 100683

Al Husaeni, D.N. (2022). Development analysis research on physics education by mapping keywords using the VOSviewer application. ASEAN Journal of Physical Education and Sport Science, 1(1), 9-18.

Alghafri, A. S. R., & Ismail, H. N. B. (2014). The effects of integrating creative and critical thinking on schools students' thinking. International Journal of Social Science and Humanity, 4(6), 518.

Azizah, E.V., Nandiyanto, A.B.D., Kurniawan, T., and Bilad, M.R. (2022). The effectiveness of using a virtual laboratory in distance learning on the measurement materials of the natural sciences of

physics for junior high school students. ASEAN Journal of Science and Engineering Education, 2(3), 207-214.

Birgili, B. (2015). Creative and critical thinking skills in problem-based learning environments.

Brookhart, S. M. (2010). How to assess higher-order thinking skills in your classroom. ASCD.

Budiman, A. Jailani.(2014). Pengembangan Instrumen Asesmen Higher Order Thinking Skill (HOTS)...(Agus Budiman, Jailani)-139. Jurnal Riset Pendidikan Matematika, 1(2), 139-151.

Cahyo, F. T. F. (2018). Hubungan antara keterampilan berpikir kritis dengan keterampilan berpikir kreatif pada beberapa model pembelajaran biologi siswa kelas XI SMA di Malang (Doctoral dissertation, Universitas Negeri Malang).

Chang, Y., Li, B. D., Chen, H. C., & Chiu, F. C. (2015). Investigating the synergy of critical thinking and creative thinking in the course of integrated activity in Taiwan. Educational Psychology, 35(3), 341-360.

Conklin, W. (2012). Higher order thinking skills to develop 21st century learners. Huntington Beach, CA: Shell Education Publishing

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

DeMars, C. (2010). Item response theory. Oxford University Press.

Du Toit, M. (Ed.). (2003). IRT from SSI: Bilog-MG, multilog, parscale, testfact. Scientific Software International.

Dwyer C P, Hogan M J and Stewart I (2014) . An integrated critical thinking framework for the 21st century Thinking Skills and Creativity 12 pp.43-52

Fatmawati, A., Zubaidah, S., & Mahanal, S. (2019, December). Critical Thinking, Creative Thinking, and Learning Achievement: How They are Related. In Journal of Physics: Conference Series (Vol. 1417, No. 1, p. 012070). IOP Publishing.

Fiandini, M., Nandiyanto, A.B.D., Al Husaeni, D.F., Al Husaeni, D.N., and Mushiban, M. (2024). How to calculate statistics for significant difference test using SPSS: Understanding students comprehension on the concept of steam engines as power plant. Indonesian Journal of Science and Technology, 9(1), 45-108.

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.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis, 8th edn., Cengage Learning EMEA, Andover. Hampshire

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory.Boston, MA : Kluwer.Inc

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

Hasudungan, A. N., & Kurniawan, Y. (2018, October). Meningkatkan Kesadaran Generasi Emas Indonesia Dalam Menghadapi Era Revolusi Industri 4.0 Melalui Inovasi Digital Platform www. indonesia2045. org. In Prosiding Seminar Nasional Multidisiplin (Vol. 1, pp. 51-58).

Hermawan, A. (2014). Mengetahui Karakteristik Peserta Didik untuk Memaksimalkan Pembelajaran. Jurnal Pendidikan Karakter, 7(1), 14-25.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The physics teacher, 30(3), 141-158.

Hidayati, N., Zubaidah, S., Suarsini, E., & Praherdhiono, H. (2019). Examining the relationship between creativity and critical thinking through integrated problem-based learning and digital mind maps. Universal Journal of Education Research, 7(9A), 171-179.

Ibrahim, A.O. (2023). Impact of blended learning method on secondary school physics students' achievement and retention in Lokoja, Nigeria. ASEAN Journal for Science Education, 2(2), 57-66.

Istiyono, E., Dwandaru, W. S. B., Erfianti, L., & Astuti, W. (2020b, January). Applying CBT in physics learning to measure students' higher order thinking skills. In Journal of Physics: Conference Series (Vol. 1440, No. 1, p. 012061). IOP Publishing.

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020a). 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.

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.

Jailani, J., Sugiman, S., Retnawati, H., Bukhori, B., Apino, E., Djidu, H., & Arifin, Z. (2018). Desain pembelajaran matematika: untuk melatihkan higher order thinking skills.

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

Kizlik, B. (2012). Measurement, assessment, and evaluation in education. Retrieved October, 10, 2015.

Krulik, S., & Rudnick, J. A. (1999). Innovative task to improve critical and creative thinking skill. Dalam L. V. Stiff & F. R Curcio (Eds.). Developing Mathematical Reasoning in Grades K-12 (pp. 138). Reston, VA: NCTM.

Lestari, D.A., Suwarma, I.R., and Suhendi, E. (2024). Feasibility analysis of the development of STEM-based physics e-book with self-regulated learning on global warming topics. Indonesian Journal of Teaching in Science, 4(1), 1-10.

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.

Luthvitasari, N., & Linuwih, S. (2012). Implementasi pembelajaran Fisika Berbasis Proyek terhadap keterampilan berpikir kritis, berpikir kreatif dan kemahiran generik sains. Journal of innovative Science education, 1(2).

Marlina, L., Tjasyono, B., & Hendayana, S. (2018, May). Improving the critical thinking skills of junior high school students on Earth and Space Science (ESS) materials. In Journal of Physics: Conference Series IOP Pubslishing (Vol. 1013, No. 1, p. 012063).

Maryanti, R., and Nandiyanto, A.B.D. (2021). Curriculum development in science education in vocational school. ASEAN Journal of Science and Engineering Education, 2(1), 151-156.

Muraki, E. (1997). PARSCALE: IRT item analysis and test scoring for rating-scale data. Scientific Software International.

Muraki, E., & Bock, R. D. (1997). Parscale 3: IRT based test scoring and item analysis for graded items and rating scales. Chicago: Scientific Software.

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.

Presseisen, B. Z. (1988). Thinking skill: meanings and models. Dalam A. L. Costa (Eds.), Developing minds: A resource book for teaching thinking (pp. 43-48). Alexandria, VA: 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).

Retnawati, H. (2008). Estimasi efisiensi relatif tes berdasarkan teori respons butir dan teori tes klasik. Disertasi. Yogyakarta: Program Pascasarjana Universitas Negeri Yogyakarta.

Retnawati, H. (2014). Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana. Yogyakarta: Nuha Medika.

Retnawati, H. (2016). Validitas reliabilitas dan karakteristik butir. Yogyakarta: Parama Publishing.

Rosina, H., Virgantina, V., Ayyash, Y., Dwiyanti, V., and Boonsong.S. (2021). Vocational education curriculum: Between vocational education and industrial needs. ASEAN Journal of Science and Engineering Education, 1(2), 105-110.

Rosmiati, R., & Satriawan, M. (2019). The ocean climate phenomenon: the challenges of earth physics lectures in Indonesia. In Journal of Physics: Conference Series IOP Publishing (Vol. 1157, No. 3, p. 032038).

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. In Journal of Physics: Conference Series (Vol. 1157, No. 3, p. 032029).

Si, C. F., & Schumacker, R. E. (2004). Ability estimation under different item parameterization and scoring models. International Journal of Testing, 4(2), 137-181.

Siburian, J., Corebima, A. D., & Saptasari, M. (2019). The correlation between critical and creative thinking skills on cognitive learning results. Eurasian Journal of Educational Research, 19(81), 99-114.

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: The Official Journal of the National Association for Research in Science Teaching, 46(6), 610-631.

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.

Susanti, E. (2019, June). Mathematical Critical Thinking and Creative Thinking Skills: How Does Their Relationship Influence Mathematical Achievement?. In Proceedings of the 2019 International Conference on Mathematics, Science and Technology Teaching and Learning (pp. 63-66).

Susilowati, N.I., Liliawati, W., and Rusdiana, D. (2023). Science process skills test instruments in the new indonesian curriculum (merdeka): Physics subject in renewable energy topic. Indonesian Journal of Teaching in Science, 3(2), 121-132.

Ülger, K. (2016). The relationship between creative thinking and critical thinking skills of students.

Ülger, K. (2020). Bloom Taksonomisi Perspektifinden Öğrencilerin Eleştirel Düşünme Ve Yaratici Düşünme Becerileri Arasindaki İlişki. International Journal of New Trends in Arts, Sports & Science Education (IJTASE), 9(2), 63-70.

Van Der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models,

and extensions. In Handbook of modern item response theory (pp. 1-28). Springer, New York, NY.

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), 020104.

APENDIX. RECAPITULATION OF MODEL FIT WITH HOTS TEST DATA

Item	GRM		PCM		GPCM2PL		GPCM3PL	
	Sig.	**Fitness**	**Sig.**	**Fitness**	**Sig.**	**Fitness**	**Sig.**	**Fitness**
1	0.367	Fit	0.133	Fit	0.422	Fit	0.000	Not Fit
2	0.006	Not Fit	0.036	Not Fit	0.003	Not Fit	0.000	Not Fit
3	0.113	Fit	0.232	Fit	0.134	Fit	0.000	Not Fit
4	0.022	Not Fit	0.033	Not Fit	0.009	Not Fit	0.000	Not Fit
5	0.007	Not Fit	0.151	Fit	0.341	Fit	0.000	Not Fit
6	0.009	Not Fit	0.236	Fit	0.440	Fit	0.000	Not Fit
7	0.000	Not Fit	0.000	Not Fit	0.000	Not Fit	0.000	Not Fit
8	0.040	Not Fit	0.311	Fit	0.373	Fit	0.000	Not Fit
9	0.000	Not Fit	0.000	Not Fit	0.001	Not Fit	0.002	Not Fit
10	0.172	Fit	0.046	Not Fit	0.100	Fit	0.001	Not Fit
11	0.001	Not Fit	0.490	Fit	0.052	Fit	0.000	Not Fit
12	0.000	Not Fit	0.115	Fit	0.009	Not Fit	0.000	Not Fit
13	0.126	Fit	0.001	Not Fit	0.004	Not Fit	0.001	Not Fit
14	0.032	Not Fit	0.034	Not Fit	0.020	Not Fit	0.000	Not Fit
15	0.000	Not Fit	0.000	Not Fit	0.003	Not Fit	0.004	Not Fit
16	0.367	Fit	0.090	Fit	0.550	Fit	0.000	Not Fit
17	0.699	Fit	0.057	Fit	0.639	Fit	0.000	Not Fit
18	0.272	Fit	0.016	Not Fit	0.476	Fit	0.000	Not Fit
19	0.128	Fit	0.137	Fit	0.806	Fit	0.000	Not Fit
20	0.026	Not Fit	0.048	Not Fit	0.314	Fit	0.000	Not Fit
21	0.112	Fit	0.347	Fit	0.831	Fit	0.000	Not Fit
22	0.074	Fit	0.384	Fit	0.118	Fit	0.000	Not Fit
23	0.000	Not Fit	0.000	Not Fit	0.087	Fit	0.028	Not Fit
24	0.118	Fit	0.210	Fit	0.060	Fit	0.000	Not Fit
25	0.023	Not Fit	0.059	Fit	0.000	Not Fit	0.000	Not Fit
26	0.000	Not Fit	0.033	Not Fit	0.284	Fit	0.000	Not Fit
27	0.015	Not Fit	0.442	Fit	0.715	Fit	0.000	Not Fit