Berkala Ilmiah Pendidikan Fisika
ISSN : 2337-604X (print)
ISSN : 2549-2764 (online)

Vol 8 No 2 2020
https://ppjp.ulm.ac.id/journal/index.php/bipf
page 103-116

The Development of HOTS Test Instrument of Thermodynamics Law for Senior High School

Yeni Rima Liana[1,2], Suharto Linuwih[1], Sulhadi[1]
[1]Physics Education Study Program, Postgraduate,
Semarang State University, Semarang, Indonesia
[2] Senior High School 2 Batang, in Batang Regency, Indonesia
yeni_rimaliana@yahoo.co.id

DOI:10.20527/bipf.v8i2.8493

Received : 21 May 2020 Accepted : 30 June 2020 Published : 30 June 2020

Abstract

The main analysis in developing assessment instruments is reliability and validity. A validity test is carried out to determine the appropriateness instrument that will be developed, both construction validity and content validity. The reliability test is to determine the level of consistency of the instrument that has been developed. This research was conducted to develop the HOTS ability to test items for high school students. The grid test instruments are arranged based on competence and HOTS indicators, which are then used to arrange items. The test instrument consisted of ten question items relating to the HOTS Thermodynamic Law problem, which included: 1) analyzing the magnitude of engine efficiency, work, heat, and internal energy changes, 2) evaluating Carnot's efficiency, and 3) creating a heat engine. The assessment of the instrument HOTS test obtained Aiken's V score in the range of 0.83 to 0.94, which is in the valid criteria. The validated instrument was piloted in 141 science grade XI student in High School 2 Batang, at Batang Regency, Central Java. The level difficulty of the Polytomous data was analyzed using the QUEST program for classical analysis and PARSCALE 4 for modern analytical theory based on the Partial Credit Model (PCM). The results of data analysis of the experimental items show that of the ten-question items, all are compatible with PCM. The reliability of the test instrument is 0.84, and the item difficulty level is in the range of 0.83 to 1.22. Information functions and Standard Error Measurement (SEM) indicate that test questions developed reliably to measure HOTS students' ability with an average category in $-1.9 < \theta < +1.7$ logit scale with SEM ± 0.5.

Keywords: HOTS test instruments; polytomous; thermodynamics law

How to cite: Liana, Y. R., Linuwih, S., & Sulhadi, S. (2020). Development of HOTS thermodynamic law test instrument for high school students in realizing learning outcomes. *Berkala Ilmiah Pendidikan Fisika*, 8(2), 103-116.

INTRODUCTION

Education is a means of improving the quality of human resources. Quality human resources prove that the level of education is quality (Baran, 2016). Quality education begins with learning

programs that are arranged systematically in accordance with the applicable curriculum design. The learning environment with structured pedagogical concepts, systematic curriculum design, and a comfortable learning atmosphere can make students transform knowledge effectively (Guney & Al, 2012).

The 2013 curriculum was designed with various improvements, including content standards, namely reducing irrelevant material, deepening and expanding material relevant to students, and enriched with students' needs to think critically and analytically following international standards. Other improvements have also been made to the assessment standards by gradually adapting international standard assessment models. Assessment of learning outcomes is expected to help students improve Higher Order Thinking Skills (HOTS) because higher-level thinking can encourage students to think broadly and deeply about the subject matter.

The Ministry of Education and Culture has begun to apply international standards, mathematics, literacy, and Natural Sciences, those requiring high reasoning power, or HOTS. HOTS ability is an important competency in the modern world, so it is a must for every student. Creativity solves problems in HOTS, consisting of above (1) ability to solve problems logically; (2) the ability to evaluate strategies used to solve problems from a variety of different perspectives; and (3) find new settlement models that are different from previous methods.

Physics is a learning activity that has the purpose of developing logical abilities and inductive and deductive analysis of students using physics concepts to solve. Physics focuses on qualitative or quantitative measurements in finding and discovering basic laws relating to phenomena and using them to develop theories. Baran (2016) states that learning physics provides the ability for someone to problem-solving in learning.

Problem-solving is the most important basic element in physics learning (Docktor, Strand, Mestre, & Ross, 2015; Yuberti, Latifah, Anugrah, Saregar, Misbah, & Jermsittiparser, 2019). Merriënboer (2013) suggested four stages of problem-solving, namely (1) studying the problems raised, (2) exploring and interpreting information with appropriate procedures, (3) looking for references that support solving problems, and (4) the process of trying to solve problems. Whereas according to Dostál (2015), analyzing problem-solving must consider several things, such as the ability to see problems, perception of problems, ability to solve problems, and problem-solving strategies.

Problem-solving strategies are very useful for solving problems in physics learning. Schoenfeld (2013) states that the process of finding a solution to a problem depends on the problem-solving strategies used. One problem-solving strategy is to involve students in communicating their ideas openly and developing HOTS through the Problem-Based Learning (PBL) learning model (Sucipto, 2017). HOTS are needed by students to improve their ability to overcome learning problems (Royantoro, Mujasam, Yusuf, & Widyaningsih, 2018). Therefore, problem-solving requires a higher level of thinking than remembering, understanding, and applying. Sambite, Mujasam, Widyaningsih, & Yusuf (2019) state that through the Project-Based Learning (PjBL) model supported by teaching aids, students can improve the HOTS ability. Students are directly involved in making tools so that it gives an imprint and profound effect.

Anderson & Krathwohl (2001) categorize the ability of the process of analyzing, evaluating, and creating, including high-level thinking. Analyzing is the ability to break things down into

smaller parts, so that deeper meaning is obtained. Analyzing the revised Bloom's taxonomy also includes the ability to organize and connect between sections so that a more comprehensive meaning is obtained. If the ability to analyze leads to a process of critical thinking so that someone can make the right decision, the person has reached the level of evaluating thinking (Setiawati, Asmira, Ariyan, Bestary, & Pudjiastuti, 2019). From the evaluation activities, someone can find weaknesses and strengths. Based on these weaknesses and strengths, finally generated ideas or new ideas or different from existing ones. When someone can produce ideas or new or different ideas, that level of thinking is called the level of thinking to create (Argaw, Haile, Ayalew, & Kuma, 2017). Someone sharp in his analysis, able to evaluate and make decisions appropriately, and always gives birth to new ideas or ideas. Therefore, the person has a great chance of solving every problem he faces (Gunawan, Harjono, Herayanti, & Husein, 2019).

Characteristics of HOTS questions: (1) transfer one concept to another; (2) processing and applying information; (3) looking for links from various different information; (4) use information to solve problems; and (5) critically examine ideas and information (Soeharto & Rosmaiyadi, 2018). Puspendik (2019) classifies cognitive levels, namely: knowledge and understanding (level 1), application (level 2), and reasoning (level 3). Reasoning level is a level of ability to think high level (HOTS) because to answer questions at level 3 students must be able to remember, understand, and apply factual, conceptual, and procedural knowledge and have high logic and reasoning to solve problems contextual (real situations that are not routine).

The level of reasoning includes the dimensions of the thought process of analyzing (C4), evaluating (C5), and creating (C6). The thought process analysis indicator (C4) demands students' ability to describe, integrate, organize, associate, diagram, compare, examine, and find implied meaning. The dimension of the process of evaluating thinking (C5) requires the ability of students to form hypotheses, prove, criticize, predict, assess, test, conclude, justify, or blame. While the dimensions of the thought process creative dimension (C6) require the ability of students to design, build, plan, produce, find, renew, perfect, strengthen, beautify, compose.

Understanding of HOTS is determined by monitoring the process, progress, and improvement of continuous learning outcomes so that an assessment is needed to measure the HOTS of students. Assessment in the world of education can use two kinds of measurement theories, namely classical theory, and modern theory. The use of classical measurement theory in Indonesia to analyze and estimate students' abilities are more desirable than modern measurement theory (Fajrianthi, Hendriani, & Septarini, 2016). However, classical measurement theory has a weakness in its use. The disadvantages of classical measurement theory include the characteristics of test items such as the level of difficulty and the power of differences that depend on students (Persichitte, 2016). Problems with classical measurement theory will have an impact on the level of ability of students that is difficult to know (Awopeju & Afolabi, 2016). Problems that arise in classical measurement theory can be solved by applying modern measurement theory, namely the approach Item Response Theory (IRT) (Baker & Frank, 2001).

IRT is a modern measurement theory that has the advantage of being able to find out the abilities and scores of students and have a more complex measurement model (Persichitte, 2016). DeMars (2013) explains that item response theory shows the relationship of ability or level trait measured using

instruments and response points with a dichotomous or polytomous scoring model. The scoring model for dichotomous grains consists of: a) 1-PL model (Logistic Parameters) which emphasizes one parameter, namely the level of difficulty of the item, b) the 2-PL model which emphasizes two parameters, namely the level of grain difficulty and power difference, and c) the 3-PL model emphasizes three parameters, namely the level of difficulty of the item, different power and pseudo guessing (Mardapi, 2012). Scoring models for polytomous items that are often used include the Graded Response Model (GRM), Modified Graded Response Model (MGRM), and Partial Credit Model (PCM) (Aybek & Demirtasli, 2017).

PCM is the development of a one-parameter logistic IRT model (1-PL) and is included in the Rasch model (Bacci, Bartolucci, & Gnaldi, 2014) PCM is a polytomous scoring model that uses several categories to analyze responses to an instrument (Masters, 2011). For example, in a vector representation test instrument developed where the process for answering, it requires several steps of completion. The PCM is the easiest and most widely applied polytomous item scoring model to analyze tests and assessments such as measuring critical thinking skills, computer-adaptive tests (CAT), measuring conceptual understanding in science and diagnosing mathematical errors. (Grunert, Raker, Murphy, & Holme, 2013) state that the PCM model is useful for knowing students' level of conceptual knowledge. The Partial Credit Model is an IRT analysis model developed to know the relationship of grain characteristics to the natural responses of students (ability or level trait). Bond & Fox (2015) states that PCM specifically combines different response levels for different items on the same test, which can combine dichotomous and polytomous items.

METHOD

This research is a kind of development research with a quantitative approach. This development research uses a 4-D development model (Define, Design, Develop, and Disseminate). The study began in October 2019 until January 2020. The development and preparation of the HOTS test instrument were conducted in October 2019 until December 2019. The trial was conducted in January 2020. The stages of test development are presented in Figure 1.

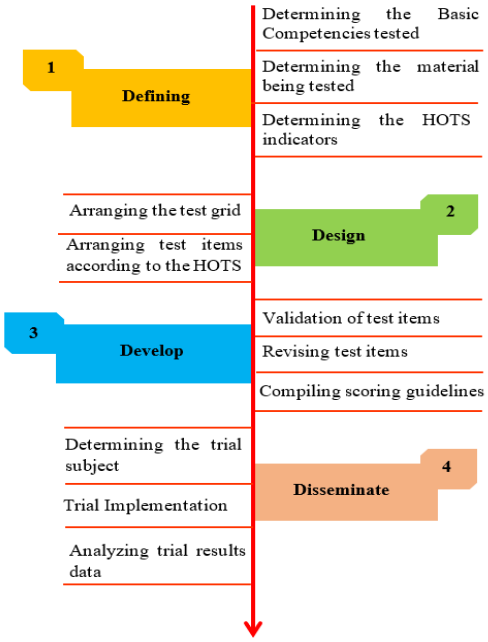

Figure 1 Steps for developing HOTS ability test instruments

The sample size used was 141 students. Bond & Fox (2015) stated that for analysis using IRT, a sample of between 30 and 300 people was used. While Reckase (2010) argues that the sample needed for analysis using 3-PL IRT, which includes the level of difficulty, power difference, and pseudo guessing, is 300 people (Haladyna & Rodriguez, 2013). The sample of this study was students of science class students of XI grade Senior High School 2 Batang, in the even semester of school

year 2019/2020 selected using the random sampling method. So that by using the PCM 1-PL model, 141 students were sufficient as subjects for empirical trials.

The technique of feasibility analysis and empirical validation of the HOTS ability test instrument uses the Aiken's V equation to calculate the content-validity coefficient as follows:

$$V = \frac{\sum s}{n(c-1)} \, , s = r - l_0$$

Where l_0 is the lowest validity rating number, c is the highest validity rating number, r is the number given by a validator, and n is the number of raters. Table 1 Product Eligibility Criteria (Azwar, 2016).

Table 1 Product feasibility criteria

Range of Scores	Categories
0,78≤ V ≤ 1,00	Valid
0,00 < V < 0,78	Invalid

The goodness of fit analysis is carried out to determine item compatibility with the PCM. The goodness of fit is analyzed by interpreting the average MNSQ INFIT value along with the standard deviation or the average INFIT value t along with the standard deviation (Adams & Khoo, 2016). If the average INFIT MNSQ approaches 1.0 and the default deviation is 0.0, or the average INFIT t approaches 0.0. The default deviation is 1.0, and then the item is said to be fit with the model. Item is said to be valid if the value if the INFIT MNSQ values in the range of values from 0.77 to 1.30 (Subali & Suyata, 2011:10-11). If converted using a standard value of t, this value is in the range of -2 to +2 (rounding from 1.96 to +1.96) with an error rate of 5% (Bond & Fox, 2015). The item is said to be good if it has a level of difficulty from -2 to +2 units of logit (Retnawati, 2011:56).

The QUEST and PARSCALE 4 programs are used to analyze the results of the trial data. Scores obtained by students were analyzed using the QUEST program to determine reliability based on internal consistency and the total information function curve. PARSCALE 4 program is used to analyze data to show parameters of item characteristics such as 1) item characteristic curve, 2) item parameter estimation, 3) estimation of student's HOTS ability, and 4) information function and standard error measurement (SEM).

RESULTS AND DISCUSSION

HOTS Ability Test Development Results

The developed test instruments amounted to 10 items in the description of thermodynamics law material. Test instruments are arranged and assembled according to HOTS indicators, which include: 1) analyzing the magnitude of efficiency, work, heat, and changes in internal energy, 2) evaluating the real efficiency and efficiency of Carnot, 3) creating a heat engine image. The developed test instrument refers to the HOTS indicator, which is part of problem-solving. Figure 2 shows an example of the HOTS capability test instrument used in this research.

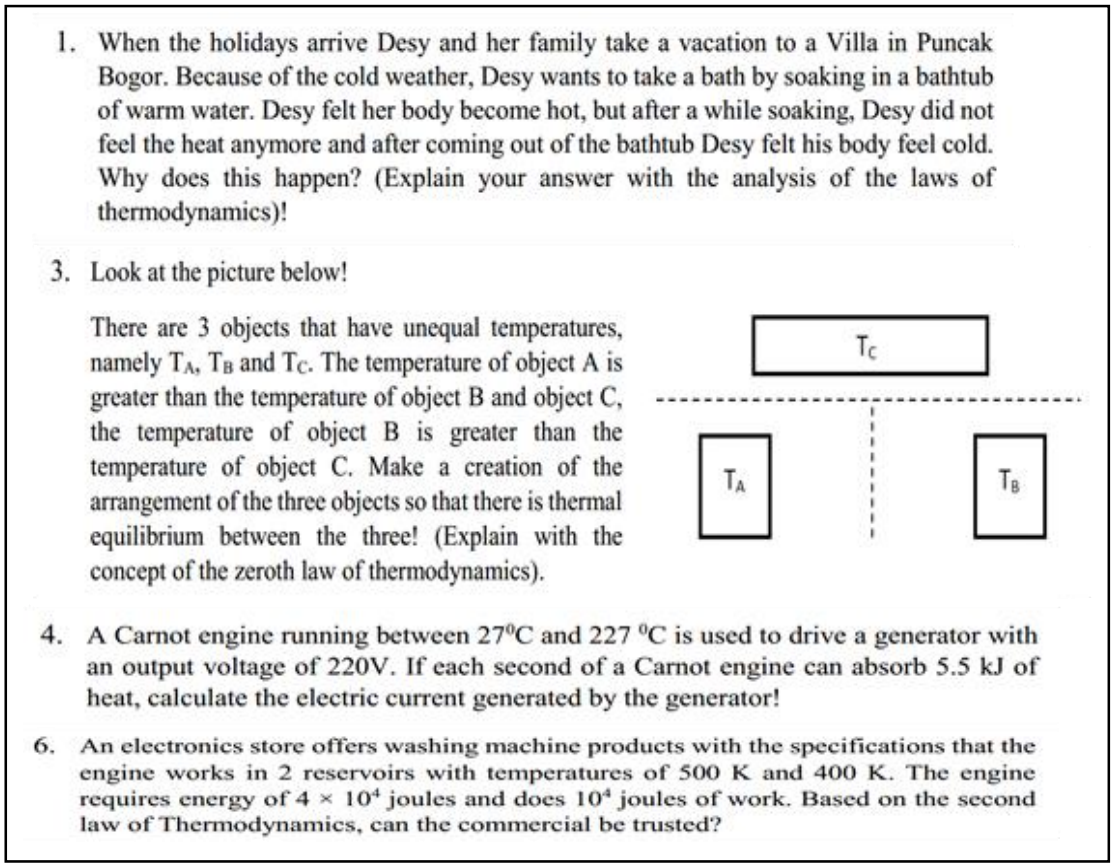

Figure 2 Four example of HOTS test instruments used

The feasibility of the test instrument was assessed based on material aspects and empirical tests (Yadiannur & Supahar, 2017). The results of the analysis material aspects using Aiken's V show that the test instruments developed are in the valid category. According to Aiken's V, this is following the validation criteria, which states that for six validators, items are declared valid if they obtain Aiken's V score V 0,78. Results of validation by expert judgment, as shown in Table 3.

Table 3 Test item validation results based on Aiken's V

HOTS Indicator	Item Number	Score of Aiken's V	Criteria
Analyzing the magnitude of efficiency, work, heat, and the change of internal energy	1, 4, 8, 10	0.89	Valid
Evaluating the real efficiency and efficiency of Carnot	5, 6, 7, 9	0.92	Valid
Creating a heat engine image	2,3	0.86	Valid

The assessment of the instrument HOTS test obtained Aiken's V score in the range of 0.83 to 0.94, which is in the valid criteria. Some improvements based on expert advice in the use of appropriate words such as "thermal balance" are replaced by "thermal equilibrium". In addition, questions need to be added that invite students to create something in problem-solving. The students' answers

from the instrument of HOTS ability on thermodynamics law material can be seen in Figure 3.

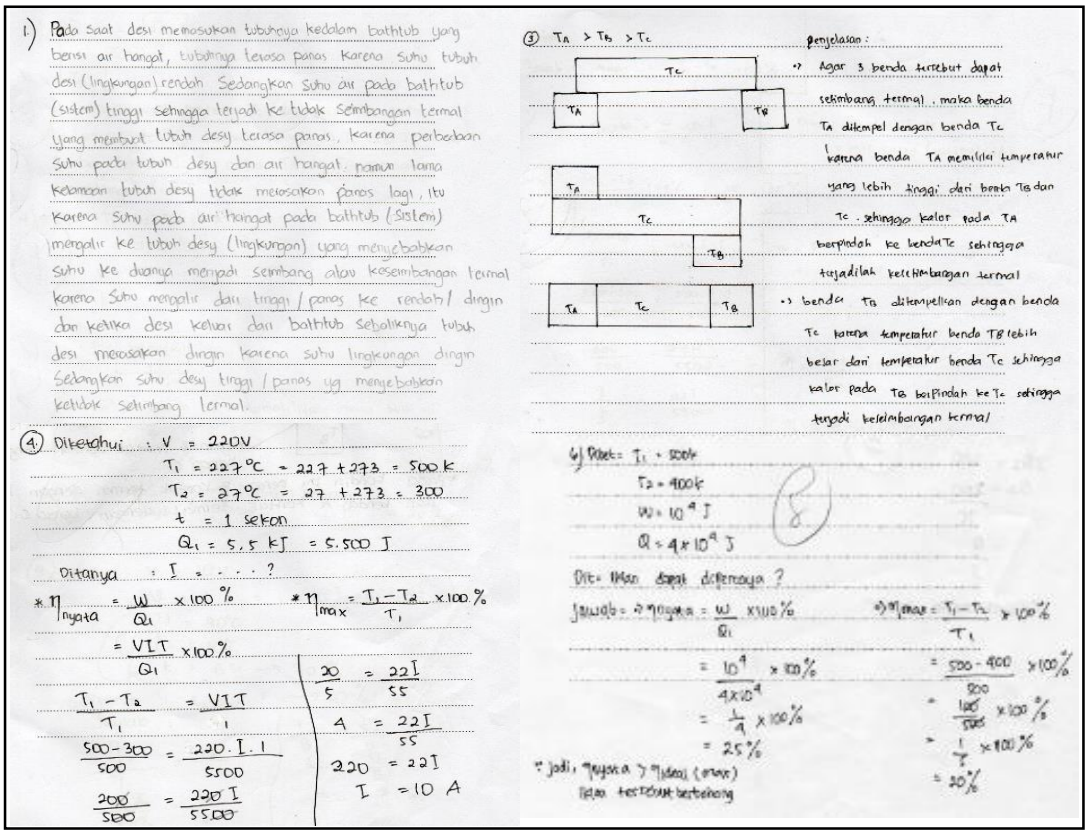

Figure 3 The students' answers from the instrument of HOTS ability

Match of the Goodness of Fit Test Item to the PCM Model

Overall, testing the goodness of fit is done by analyzing the results of the trial test questions using the Quest program. The goodness of fit is tested according to the rules developed by (Adams & Khoo, 2016). They look at the average value of INFIT MNSQ and its standard deviation or by observing the average value of INFIT t and its standard deviation. The test instrument is said to be fit with the 1-PL PCM model if the average INFIT MNSQ value is around 1.0 and the standard deviation is 0.0 or the INFIT average value is around 0.0, and the default deviation is 1.0. Table 4 shows items and test estimates from the HOTS ability test instrument.

Table 4 Item estimation and test of the test instrument

Description	Item Estimation	Test Estimates
Average value and standar deviation	0.00 ± 0.40	0.07 ± 1.44
Reliability	0.84	0.80
Average value and standard deviation of INFIT MNSQ	0.99 ± 0.14	0.98 ± 0.43
Average value and standard deviation of OUTFIT MNSQ	0.99 ± 0.14	0.99 ± 0.46
Average value and standard deviation of INFIT t	-0.04 ± 1.13	0.06 ± 1.20
Average value and standard deviation of OUTFIT t	-0.25 ± 0.23	0.17 ± 1.17

Testing of the goodness of fit for each item follows the rules developed by Adams & Khoo (2016) by looking at the INFIT MNSQ value of each item based on the output of the QUEST program. The item is said to be valid if the value if the MNSQ INFIT value ranges from 0.77 to 1.30. Besides, items are also declared fit to the model if the INFIT t value is in the range of -2 to +2. Table 5 shows the INFIT MNSQ and INFIT values for each item.

Table 5 Distribution of INFIT MNSQ and INFIT t each test item

Item Number	INFIT MNSQ	OUTFIT MNSQ	INFIT t	OUTFIT t
1	1.20	1.20	1.20	0.60
2	0.87	-0.06	-0.06	-0.10
3	0.83	-0.50	-1.50	-0.40
4	1.22	0.30	0.30	0.20
5	0.96	1.80	1.80	1.60
6	1.09	-0.04	-0.04	-0.10
7	0.88	-1.80	-1.80	0.30
8	0.88	-0.30	-0.30	0.10
9	0.97	0.70	0.70	0.30
10	1.04	-1.00	-1.00	1.10
Average	0.99	0.03	-0,50	0,40

Table 5 shows that the HOTS ability test items developed to have a range of INFIT MNSQ values from 0.83 to 1.22. This value indicates that all items have MNSQ INFIT values within the range of acceptance of the goodness of fit, so that it is concluded that all test items valid and match the partial credit model (PCM).

Reliability

Reliability obtained based on analysis using the QUEST program is 0.84. The reliability value obtained has a high category. This reliability value indicates that the vector representation ability test instrument developed is qualified as a good instrument.

Item Characteristic Curve

Item characteristics are indicated by the item characteristic curve (ICC). Analysis to find out the PARSCALE 4 program used the ICC. The analysis carried out was obtained as many as ten items characteristic curves. Figure 4 presents ICC item number 1.

The ICC chart in Figure 4 shows students' opportunity to answer test items based on their abilities. Opportunities for students working on item number 1 are as follows: 1) category 1 is, 2) category 2 is, 3) category 3 is, 4) category 4 is, 5) category 5 is.

ICC for item number 1 contains information as follows: 1) category 1 mostly obtained by students with ability -3.50 logit scale. 2) category 2 is mostly obtained by students who have the ability of -1.50 logit scale. 3) category 3 is mostly obtained by students who have a capability of 0.20 logit scale. 4) category 4 mostly obtained by students who have the ability of 1.00 logit scale. 5) category 5 is mostly obtained by students who have the ability of 3.50 logit scale.

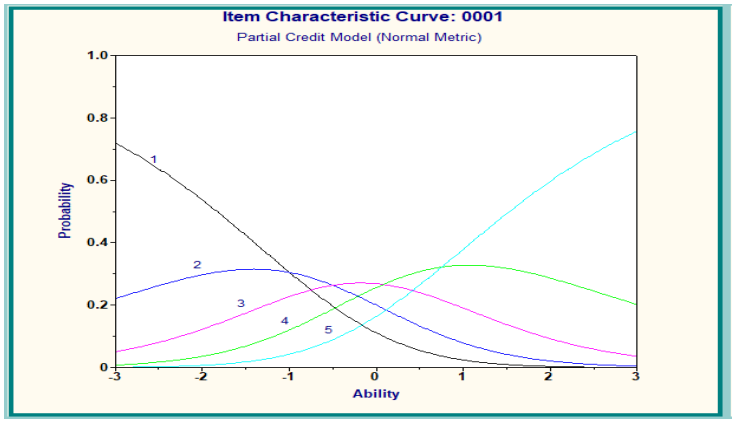

Figure 4 Characteristic curve number 1

Item Parameter Estimation

According to the PCM model, the estimated parameters of vector representation ability test items are indicated by different difficulty levels for each item. Table 6 shows a summary of parameter estimates analyzed using the PARSCALE 4 program.

Table 6 Test item parameter estimation

PARAMETER	MEAN	STN DEV	N
SLOPE	0.589	0.000	10
LOG (SLOPE)	-0.529	0.000	10
THRESHOLD	-0.114	0.398	10
GUESSING	0.000	0.000	0

The power estimation of different items is indicated by the SLOPE parameter, which has an average value of 0.589. The item's level of difficulty is indicated by the THRESHOLD parameter, which has an average value of -0.114. The pseudo guessing parameter is shown by the GUESSING parameter, which has a value of 0.000. Partial Credit Model (PCM) 1-PL refers to one parameter: the difficulty level of an item. Table 6 shows the difficulty level of each HOTS ability item for each score category in PCM

Table 7 Level of difficulty test item thermodynamics law ability

Item No	Difficulty	Stage Difficulty				
		Category 1	Category 2	Category 3	Category 4	Category 5
1	-0.165	-0.63	0.50	0,88	-0.68	-0.07
2	-0.051	-0.85	0.28	-0.69	0.13	1.13
3	-0.672	-0.89	-0.10	0.35	0.24	0.40
4	-0.342	-1.27	1.17	0.28	-0.12	-0.07
5	0.508	-0.88	-0.32	0.99	0.25	-0.03
6	-0.91	-1.62	-1.07	0.13	1.02	1.53
7	0.51	-2.82	0.32	-0.04	1.09	1.45
8	-0.32	-2.37	-0.21	-0.40	1.52	1.46
9	0.62	-1.46	0.49	0.14	0.39	0.40
10	0.04	-0.43	-1.12	-0.57	1.23	0.88

Table 7 shows that PCM measures students' ability to work on test items based on the steps taken by dividing them into several categories. Each category has different difficulty levels for each item. Different difficulty levels indicate the estimated parameters of HOTS ability test items according to the PCM model for each item. Partial Credit Model (PCM) 1-PL refers to one parameter, namely the difficulty level of an item. The study's findings show that the difficulty level of each item vector representation ability is divided for each score category in PCM. PCM measures students' ability to work on test items based on the steps taken by dividing them into several categories. Each category has different difficulty levels for each item. This result agrees with the research of Grunert et al. (2013), which states that the use of partial credit which is divided into several categories. Each category has different difficulty levels for each item. This result agrees with the research of Grunert et al. (2013), which states that the use of partial credit, which is divided into several categories gives a significant impact on the item being tested. The results of the research in Table 6 on the difficulty column show each item's difficulty level. The difficulty value or the difficulty of the item is in the range of -2 to +2. This value is in accordance with the opinion of Bond & Fox (2015), which states that the level of difficulty for items with good categories is in the range of -2 to +2 (rounding from -1.96 to +1.96) with an error rate of 5%. Bond & Fox (2015) opinion is supported by Hambleton, R.K Swaminathan (1985), which shows that the item is said to be good if it has a difficulty level from -2 up to +2 logit scale.

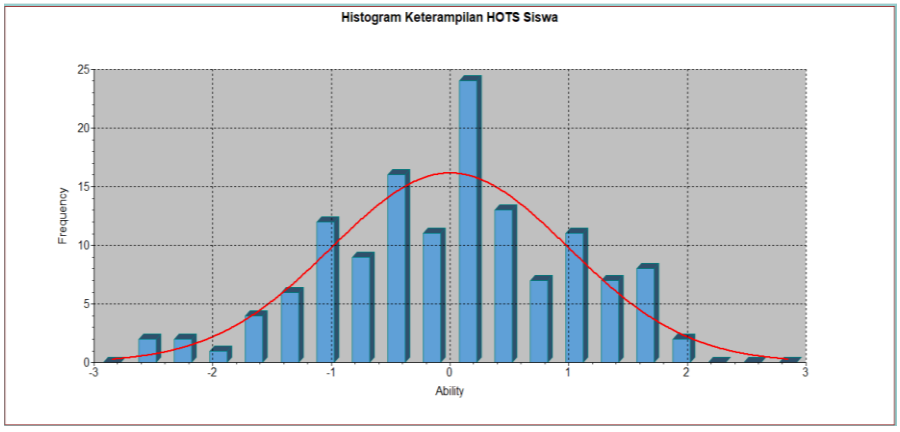

Figure 5 Histogram estimated HOTS ability

Estimating the level of ability of students is shown by the histogram. Figure 5 shows that the HOTS ability of students follows the normal curve.

Table 8 Student vector representation capability category

Sample	Ability (Logit Scale)	Interpretation
3	+2.00 to +3.00	Very high
31	+1.00 to +2.00	High
82	-1.00 to +1.00	Medium
21	-2.00 to -1.00	Low
4	-3.00 to -2.00	Very low

Table 8 shows that there are 2.83% of students who have very low HOTS abilities. There are 2.13% of students who have very high HOTS ability, 22.00% have high HOTS ability, 58.16% have medium HOTS ability, and 14.89% have low HOTS ability. The results of the study in Table 7 show that the ability of vector representation of students is in the average to very high category. This proves that the HOTS ability test item developed can measure the level of students' ability.

The results of this study agree with the Master's statement in Linden (2016) which explains that PCM is the easiest and most widely applied polytomous item scoring model to analyze tests and assessments such as measuring critical thinking skills, Computer Adaptive Test (CAT), measuring conceptual understanding in science and diagnose mathematical errors. DeMars (2013) also explains that the use of item response theory in assessment can show the relationship between ability or level trait measured using instruments and response items with dichotomous or polytomous scoring models. The same thing is shown by Aybek & Demirtasli (2017) that IRT can show the relationship between ability measured using instruments with polytomous scoring models. Based on findings, IRT approach shows that HOTS students' ability in physics learning has a medium category. But, students are almost never drilled to apply HOTS tests to solve problems in physics learning. So, the assessment HOTS needs to be developed.

Information Function and Standard Error Measurement (SEM)

Information functions and standard error measurement (SEM) were obtained based on analysis using the PARSCALE 4 program. Figure 6 shows a graph of total functions and SEM. The analysis results obtained intersection of information function lines and SEM lines at the point -1.9 up to +1.7 logit scale.

These results indicate that the student's HOTS ability is classified as medium. This value indicates that the HOTS ability test instrument was developed reliably when tested on students with a medium ability that is logit scale with SEM. This shows that the problem given is good reliability so that it can measure what it wants to measure.

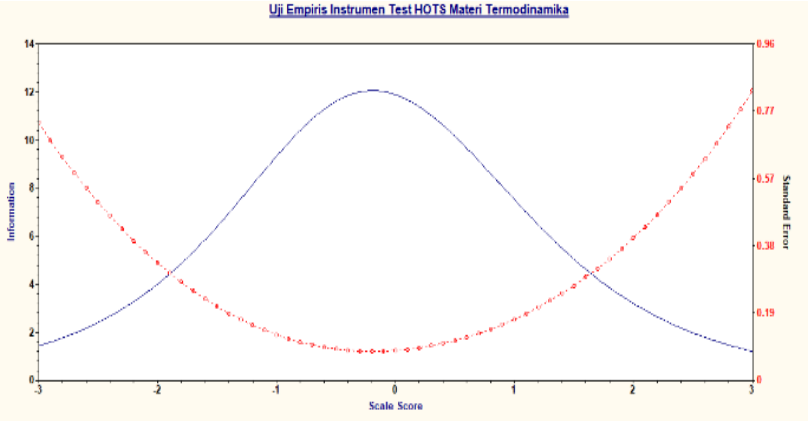

Figure 6. Information Function and Standard Error

CONCLUSION

Based on the explanation presented above, it can be concluded that: 1) The HOTS ability test items developed have a range of INFIT MNSQ values from 0.83 to 1.22. This value indicates that all test items valid and match the partial credit model (PCM). ; 2) The reliability of the test instrument is 0.84 that had a high category. ; 3) the item difficulty level is in the range 0.83 to 1.22 shows that the item is said to be good; and 4) Information functions and SEM indicate that test questions were developed reliably to measure the ability of HOTS students with an average category in $-1.9 < \theta < 1.7$ logit scale with SEM ± 0.5 t should be on the characteristics of good instruments, so the conclusion is whether the instruments are made, valid and reliable.

REFERENCES

Adams, R. J., & Khoo, S. (2016). *Quest: The ınteractive test analysis system version 2.3*. Victoria: The Australian Council for Educational Research.

Anderson, L., & Krathwohl, D. (2001). *A taxonomy for learning, tea¬ching and assessing*. New York: Longman.

Argaw, A. S., Haile, B. B., Ayalew, B. T., & Kuma, S. G. (2017). The effect of problem based learning (PBL) instruction on students' motivation and problem solving skills of physics. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(3), 857–871. https://doi.org/10.12973/eurasia.2017.00647a

Awopeju, & Afolabi. (2016). Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination. *European Scientific Journal, ESJ*, *12*(28), 263. https://doi.org/10.19044/esj.2016.v12n28p263

Aybek, E. C., & Demirtasli, R. N. (2017). Computerized Adaptive Test (CAT) Applications and Item Response Theory Models for Polytomous Items. *International Journal of Research in Education and Science*, 475–475. https://doi.org/10.21890/ijres.327907

Azwar, S. (2016). *Reliabilitas dan Validitas Edisi 4*. Yogyakarta: Pustaka Pelajar.

Bacci, S., Bartolucci, F., & Gnaldi, M. (2014). A class of multidimensional latent class IRT models for ordinal polytomous item responses. *Communications in Statistics - Theory and Methods*, *43*(4), 787–800. https://doi.org/10.1080/03610926.2013.827718

Baker, & Frank. (2001). *The basic of Item Response Theory*. USA: USA: ERIC Clearinghouse on Assessment and Evaluation.

Baran, M. (2016). An Analysis on High School Students' Perceptions of Physics Courses in Terms of Gender (A Sample from Turkey. *Journal of Education and Training Studies*, *4*(3), 150–160. https://doi.org/10.11114/jets.v4i3.1243

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model. In *Fundamental Measurement in the Human Sciences* (3rd ed.). New York: Routledge.

DeMars, C. (2013). Item Response Theory. *The Canadian Journal of Program Evaluation*, *27*(1), 126–128.

Docktor, J. L., Strand, N. E., Mestre, J.

P., & Ross, B. H. (2015). Conceptual problem solving in high school physics. *Physical Review Special Topics - Physics Education Research*, *11*(2), 1–13. https://doi.org/10.1103/PhysRevSTPER.11.020106

Dostál, J. (2015). Theory of Problem Solving. *Procedia - Social and Behavioral Sciences*, *174*, 2798–2805. https://doi.org/10.1016/j.sbspro.2015.01.970

Fajrianthi, F., Hendriani, W., & Septarini, B. G. (2016). Pengembangan Tes Berpikir Kritis Dengan Pendekatan Item Response Theory. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(1), 45. https://doi.org/10.21831/pep.v20i1.6304

Grunert, M. L., Raker, J. R., Murphy, K. L., & Holme, T. A. (2013). Polytomous versus dichotomous scoring on multiple-choice examinations: Development of a rubric for rating partial credit. *Journal of Chemical Education*, *90*(10), 1310–1315. https://doi.org/10.1021/ed400247d

Gunawan, G., Harjono, A., Herayanti, L., & Husein, S. (2019). Problem-based learning approach with supported interactive multimedia in physics course: Its effects on critical thinking disposition. *Journal for the Education of Gifted Young Scientists*, *7*(4), 1075–1089. https://doi.org/10.17478/jegys.627162

Guney, A., & Al, S. (2012). Effective Learning Environments in Relation to Different Learning Theories. *Procedia - Social and Behavioral Sciences*, *46*, 2334–2338. https://doi.org/10.1016/j.sbspro.2012.05.480

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items. Developing and Validating Test Items*. https://doi.org/10.4324/9780203850381

Mardapi, D. (2012). *Pengukuran penilaian dan evaluasi pendidikan. Nuha Medika*. Yogyakarta: Nuha Medika.

Merriënboer, J. J. G. va. (2013). Perspectives on problem solving and instruction. *Computers and Education*, *64*, 153–160. https://doi.org/10.1016/j.compedu.2012.11.025

Persichitte, K. A. (2016). *Educational Technology to Improve Quality and Access on a Global Scale. In Educational Technology World Conference (ETWC 2016) Editors: Persichitte, Kay, Suparman, Atwi, Spector, Michael (Eds.)*.

Puspendik. (2019). *Buku Penilaian Berorientasi Higher Order Thinking Skills*. Jakarta: Kementerian Pendidikan dan Kebudayaan.

Reckase, M. D. (2010). *Response theory*. https://doi.org/10.1007/bfb0044597

Retnawati, H. (2011). Mengestimasi Kemampuan Peserta Tes Uraian Matematika Dengan Penskoran Politomus dengan Generalized Partial Credit Model. *Seminar Nasional Penelitian, Pendidikan Dan Penerapan MIPA*, 53–62. Retrieved from https://doi.org/eprints.uny.ac.id:7173

Royantoro, F., Mujasam, M., Yusuf, I., & Widyaningsih, S. W. (2018). Pengaruh model problem based learning terhadap higher order thinking skills peserta didik. *Berkala Ilmiah Pendidikan Fisika*, *6*(3), 371–382.

Sambite, F. C., Mujasam, M., Widyaningsih, S. W., & Yusuf, I. (2019). Penerapan project based learning berbasis alat peraga sederhana untuk meningkatkan HOTS peserta didik. *Berkala Ilmiah Pendidikan Fisika*, *7*(2), 141–147.

Schoenfeld, A. H. (2013). Reflections on Problem Solving Theory and Practice. *The Mathematics Enthusiast*, *10*(1), 9–34.

Setiawati, W., Asmira, O., Ariyan, Y., Bestary, R., & Pudjiastuti, A. (2019). Buku Penilaian Berorientasi Higer Order Thinkings Skills (HOTS. In *Dirjen GTK*. Jakarta: Kemdikbud.

Subali, B., & Suyata, P. (2011). *Panduan Analisis Data Pengukuran Pendidikan untuk Memperoleh Bukti Empirik Kesahihan Menggunakan Program Quest*. Yogyakarta: Lembaga Penelitian dan Pengabdian Masyarakat UNY.

Sucipto, S. (2017). Pengembangan Ketrampilan Berpikir Tingkat Tinggi dengan Menggunakan Strategi Metakognitif Model Pembelajaran Problem Based Learning. *Jurnal Pendidikan (Teori Dan Praktik*, *2*(1), 77. https://doi.org/10.26740/jp.v2n1.p77-85

Yadiannur, M., & Supahar. (2017). Mobile Learning Based Worked Example in Electric Circuit (WEIEC) Application to Improve the High School Students' Electric Circuits Interpretation Ability. *International Journal of Environmental and Science Education*, *12*(3), 539–558. https://doi.org/10.12973/ijese.2017.1246

Yuberti, Y., Latifah, S., Anugrah, A., Saregar, A., Misbah, M., & Jermsittiparsert, K. (2019). Approaching problem-solving skills of momentum and impulse phenomena using context and problem-based learning. *European Journal of Educational Research*. https://doi.org/10.12973/eu-jer.8.4.1217