Development of Instrument Assessment for Learning the Polytomous Response Models to Train Higher Order Thinking Skills (HOTS)

Development of Instrument Assessment for Learning the Polytomous Response Models to Train Higher Order Thinking Skills (HOTS)

A Prasetya[1]*, U Rosidin[1], K Herlina[1]

[1]Physics Education of Graduate Program, Lampung University, Jl. Sumantri Brojonegoro No. 01, Lampung, Indonesia.

*andika_prasetya79@yahoo.com

Abstract. This study aims to determine the validity, practicality, and effectiveness of the assessment for learning instrument polytomous response model to train higher order thinking skills. This research method is based on research and development (R & D) adapted from Gall, et al. with development steps, namely research and information collection, planning, initial product development, limited trials, initial product revisions, field trials, and final product revisions. Data sources in this development are students and high school class X teachers in Lampung province. The instruments used are questionnaires and test instruments. Data analysis techniques using validity test, practicality test, and N-gain test. Research shows that the assessment for learning instrument polytomous response model to train higher order thinking skills as a reliable value of 0,91; the instrument's validity value is 0,82; the practicality value of the instrument is 0,82, and the instrument effectiveness value is 0,74. It can be concluded that the product developed is an assessment for learning an instrument of polytomous response model which has very high validity, very high practicality, and high effectiveness in training higher order thinking skills.

1. Introduction

The 2013 curriculum emphasizes that students are able to observe, ask, reason, and communicate what they have gained after receiving a lesson. In addition, the 2013 curriculum views that knowledge cannot be transferred directly from teacher to student. Students at the high school level in particular, not only have low-level thinking skills (lower order thinking, LOT) but must arrive at the ability to think higher (higher order thinking, HOT) in learning. The aim of the 2013 Curriculum is to be expected to improve the quality of education in Indonesia. Besides the curriculum, things that need to be considered in improving the quality of education are the assessment process in the classroom [1]. Forms of assessment in class that is often used in education include objective forms and descriptions. This form of assessment is used to determine students' ability to understand a learning material. The assessment of the objective form usually uses a dichotomy model scoring, it is correct to score 1 and incorrectly cast 0 [2].

The ability to think high-level in physics material includes the ability to analyze, evaluate, and create indispensable for the advancement of higher physics learning [3]. One that can help to see the development of HOTS is the need to assess students' HOTS [4]. High-level thinking ability is the widespread use of the mind to find new challenges [5]. Higher-order thinking skills require a person to apply new information or knowledge he already has and manipulate information to reach possible answers in new situations. Students are able to associate classroom learning with what happens in life [6]. The dimension of knowledge, which consists of 4 categories of knowledge [7]. The dimensions of knowledge in question are as follows: factual knowledge is knowledge in the form of separate pieces of information or basic elements that exist in a

particular scientific discipline. Conceptual knowledge is the knowledge that shows the interrelations between the basic elements in a larger structure and all of them function together. Procedural knowledge is procedural knowledge is knowledge about how to do something that can be in the form of activities or procedures. Metacognition knowledge is knowledge or activity that regulates cognition [8].

Written tests are a set of questions or assignments in written form that are planned to measure or obtain information about the ability of test participants [9]. Written tests require a response from test participants that can be used as a representation of their abilities [10]. The multiple choice item consists of the subject matter and the choice of answers [11]. Answer choices have the right or the most appropriate choice of answers and the other answer choices are the deceiver [12].

Assessment is the most important thing that can be done in helping students learn [13]. Assessment is used to determine whether students have gained enough knowledge, skills, etc. Assessment should be able to measure the ability of students in mastering the material, but can also be an assessment for learning. Shifting the assessment of learning outcomes towards assessment for learning or often referred to as assessment for learning [14]. Using assessment for learning to support learning improvement, they will achieve the goals of learning. The right strategy of teachers and students in developing learning can be to support learning. Assessment for learning is proven to improve learning [15]. Assessment of good assessment for learning is an assessment which is carried out by anyone who will produce the same value, this can be achieved if there is an assessment rubric.

Assessment for learning in the opinion of experts is to serve students to learn more so that if the learning objectives have not been achieved, the next learning process can be immediately changed in order to achieve the goals of learning [16]. Assessment for learning is more than just assessing students and providing teachers with assessment results to revise instructional learning [17]. Assessment for learning carried out by the teacher, the assessment results can be used to modify the learning process [18]. Assessment for learning is generally recognized that the use leads to a higher rate of process learning[19]. Teachers should use assessment to plan lessons, identify students' needs in learning and re-teach material that is not well understood by students [20].

Instrument assessment for learning the polytomous model is an instrument to assess the learning process. This instrument is in the form of polytomous or multiple reasoned options where each step in answering has a certain value that does not have to be the same. This assessment is better than the assessment that is often used by teachers. Because the assessment of the teacher's instruments every step has the same value. An alternative approach that can be used is the item response approach for polytomous scoring. Early development of polytomous item response theory, the more well-known model is the expansion of Rasch models called the partial credit model (PCM). In items that are stored in a tiered category, the value of difficulty in each step does not need to be sequential, a step can be more difficult or easier than the previous stage [21]. The polytomous model has two model parameters, namely the level of difficulty and different parameters [22]. In terms of constructs, PCM questions need more complex skills in compiling them, especially with regard to selecting answers to each item. Determination of alternative answers in the form of objective tests is very important related to the level of opportunity to take the test to answer correctly by guessing. An alternative form of test that can be used is the application of the PCM-shaped polytomous model scoring [23]. The dichotomous scoring model cannot be used to find errors made by students, because all wrong choices are given a score of 0 and the correct answer choice is given a score of 1. Therefore, it is necessary to score a polytomous response model.

2. Methods and Materials

This research method is based on research and development research (R & D) design [24], using 7 steps of development consisting of research and information gathering, planning, initial product development, limited trials, product revision start, field trial, and final product revision. The first phase yang do is research and collection of information. Needs analysis was carried out on 30 students and 5 physics teachers in Pringsewu District, South Lampung Regency, and Pesawaran District. Preliminary research data was collected using a questionnaire. Preliminary research is carried out to obtain the real conditions of

instruments used by teachers in schools [25]. This preliminary study obtains quantitative data descriptively, then the data in the analysis. After analyzing the data, the design of the assessment for learning model of the polytomous response model was carried out starting from determining the indicators to be used to designing the products developed. Furthermore, at the stage of the validation test, an expert test was conducted to determine the validity of the instrument by using three validators. Expert validation test uses expert test questionnaire by giving scores 5, 4, 3, 2, and 1 with answer choices according to the question content, namely: "Very valid", "Valid", "Fairly Valid", "Less Valid" and "No Valid ". The results of the assessment are then searched for [26]by using the formula:

$$\text{Score Average} = \frac{Total\ Score}{The\ Total\ of\ Experts}$$

After obtaining the average, then converted to an assessment statement. The conversion of scores [27]to an assessment statement is presented in Table 1.

Tabel 1 Criteria Score

Average Score	**Classification**
0,80 ≤ rxy ≤ 1,00	Very High
0,60 ≤ rxy< 0,80	High
0,40 ≤ rxy< 0,60	Medium
0,20 ≤ rxy< 0,40	Low
0,00 ≤ rxy< 0,20	Very Low

The next stage is the product testing phase, this stage is carried out on 30 students to find out the value of the reliability of the results of development assessment instruments. Analysis of the reliable value of instruments using ministep 4.0.1 software. After obtaining a reliable value of the instrument, it is then converted to an assessment statement. Conversion becomes an assessment statement using classification, namely: 0,80 <rxy with the classification "Very Good"; 0,70 ≤ rxy<0,80 rxy "Good"; 0,60 ≤ rxy<0,70 with the classification "Enough"; 0,50 ≤rxy<0,60 with the classification "Ugly"; rxy<0,50 with the "Bad" classification.

The practicality testing phase of the instrument using the assessment for the learning instrument response polytomous model was tested to three high school teachers. Practicality test uses practicality test questionnaire by giving scores 4, 3, 2, and 1 with answer choices according to the question content, namely: "Very Practical", "Practical", "Less Practical", "Invalid". The results of the assessment are then searched for[28] by using the formula:

$$\text{Score Average} = \frac{Total\ Score}{The\ Total\ of\ Experts}$$

After obtaining the average, then converted to an assessment statement. The conversion of scores into an assessment statement is presented in Table 1.

The field testing phase using assessment for learning instruments is a polytomous model of response in learning activities. The data collected came from the students' pretest and posttest scores. This value is used to determine the effectiveness of the product in training students' higher order thinking skills. Analysis of the results of the pretest and posttest to test the effectiveness of the product used the n-gain score. The n-gain score[29] is obtained from the following formula:

$$(g) = normalized\ gain = \frac{post\ test - pre\ test}{skor\ maksimum - pre\ test}$$

The results of the gain calculation are then interpreted with the average normalized gain using

classification, namely: (g) ≥ 0.70 with the "High" classification; 0,30≤ (g)> 0,70 with the classification "Medium"; and (g) <0.30 with a "Low" classification. Knowing that there is an increase between pretest and posttest instruments is said to be effective in training higher order thinking skills. The basis of decision making is based on the classification of the n-gain value of the pretest and posttest analysis.

3. Results and Discussion

The results of this research and development are an assessment for learning instrument polytomous response model to train higher order thinking skills. The results of each stage of the development procedure are as follows: the research stage and information gathering. This stage it is used to find out the instruments used by the teacher, whether referring to high-level thinking indicators or not because in the 2013 curriculum students are required to be able to think high. Analysis of potential and problems was carried out in five physics teachers in several schools namely; SMAN 1 Adiluwih, Pringsewu Regency, MA Al-hidayah, South Lampung Regency, and MA Al-Hidayat Gerning, Pesawaran District, as well as 30 students of class X1 MIA 1 at SMAN 1 Adiluwih, Pringsewu Regency. After the researchers conducted preliminary research to find out information in schools the researchers also analyzed the results of the preliminary research questionnaire. Based on the needs analysis, it is known that 20% of teachers stated that they had an assessment for learning instrument to measure HOTS students, 20% of teachers and 9.67% of students stated that teachers used assessment for learning instruments of polytomous response models to measure HOTS students, 0% of teachers stated that they were willing to assessment for learning instrument polytomous response model to measure HOTS students, as well as 80% teachers and 79.17% students aim when developed assessment for learning instrument polytomous response model to measure HOTS students. Based on a literature review conducted by researchers, the instruments available in schools are only examples of conventional assessment instruments that do not fit the polytomous response model to measure higher-order thinking skills in physics learning. Available instruments are dichotomy and description instruments. so far there has been no development of an assessment for learning instrument for polytomous response models to train students' higher order thinking skills (HOTS) in physics learning. The instruments that have been developed are assessment for learning instruments in physics learning in simple harmonic motion material.

Planning phase, at this stage researchers design development product designs [30]. The design of product development consists of three parts. The first part consists of the front cover, foreword, table of contents. In the middle section consists of grids, items, and scoring rubrics. The final part consists of recapitulation of values, bibliography and back cover. The instrument developed consisted of 30 questions about the polytomous response model with 15 for the pretest and 15 for the posttest. The instrument to be developed is accompanied by an assessment rubric so that anyone who uses it can objectively assess it not subjectively. The instruments developed were examined related to construct validity, substance validity and language validity. The instrument validation test developed was tested to 3 experts in the constructive, substance, and language aspects. Retrieval of validator test data by using expert test questionnaires related to construction, substance, and language [31]. The results of the instrument validation test were obtained by using the expert validation expert test questionnaire which was tested to validator 1, validator 2, and validator 3. After validating the test, the researcher analyzed the validation questionnaire results. The instrument test results data can be seen in Table 2.

Table 2 Instrument Validation Value

Aspect	Validator			Average	Classification
	1	2	3		
Construction	77,14%	80,00%	88,57%	81,90%	Very High
Contains	72,00%	84,00%	76,00%	77,33%	High
Language	80,00%	86,67%	86,67%	84,45%	Very High

The validation test questionnaire contains 7 aspects of construction. The average total score of the three

validators regarding the fulfillment of the construction aspects of the instruments that have been developed is 81.90% which means very high. The validation test questionnaire contains 5 aspects regarding the substance. The average total score of the three validators regarding the fulfillment of the substantial aspects of the instruments that have been developed is 77,33% which means high. The validation test questionnaire contains 3 aspects of the language. The average total score of the three validators regarding the fulfillment of the language aspects of the instruments that have been developed is 84,45% which means very high. The value of each aspect is more than 60%, means fulfilling the validity criteria of the instrument. This, the instrument can be said to be valid and can be used in training high-level thinking skills [32].

Limited trial phase, at this stage the development instrument was limited to 30 students of class XI. In this limited trial, it was found that the reliability value of 15 Pretest Questions can be seen in Figure 5 and the reliability value of 15 Pretest Questions can be seen in Figure 1. The results of the reliability analysis of the questions revealed that the Cronbach Alpha score was 0,92. This shows the reliability of the Pretest question in the excellent category. The value of Reliability Measured Item of 0,90 means that the quality of the items in the reliability aspect instrument is good. The value of Reliability Measured Person is 0,92, it can be concluded that the consistency of answers from students is good reliability. Separation Measured Item value of 3,06 means that the problem consists of 4 questions. The group of questions is a group of questions that are very easy, groups of questions are easy, groups of questions are difficult, and groups of questions are very difficult. Separation Measured Person value of 3,33 means that students consist of 4 groups of students. The group of students is a group of students with low abilities, groups of students with less ability, groups of students with sufficient abilities, and groups of students with good abilities. The Mean Measured Person value of -0,87 shows that the average student ability tendency is smaller than the difficulty level of the pretest problem [33].

SUMMARY OF 30 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	16.5	15.0	-.87	.45	.96	-.1	.96	.0
P.SD	9.0	.0	1.65	.09	.27	.8	.31	.7
S.SD	9.2	.0	1.68	.09	.28	.8	.31	.7
MAX.	31.0	15.0	1.42	.63	1.61	1.5	1.62	1.4
MIN.	4.0	15.0	-3.63	.35	.45	-1.6	.42	-1.2

REAL RMSE	.47	TRUE SD	1.58	SEPARATION	3.33	Person RELIABILITY	.92
MODEL RMSE	.45	TRUE SD	1.59	SEPARATION	3.49	Person RELIABILITY	.92
S.E. OF Person MEAN = .31							

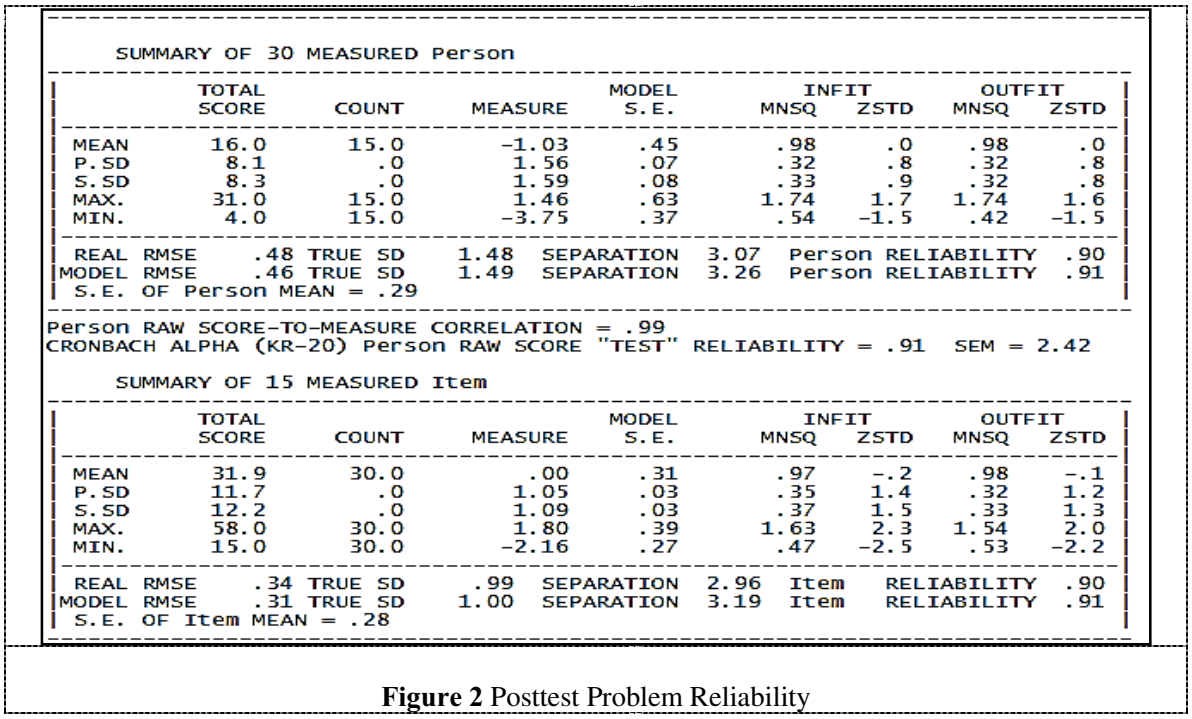
Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .92 SEM = 2.54

SUMMARY OF 15 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	33.1	30.0	.00	.30	.93	-.3	.96	-.1
P.SD	12.9	.0	1.05	.04	.37	1.4	.32	1.2
S.SD	13.3	.0	1.09	.04	.38	1.5	.33	1.2
MAX.	58.0	30.0	1.82	.40	1.64	2.3	1.49	1.8
MIN.	15.0	30.0	-1.96	.26	.49	-2.5	.49	-1.8

REAL RMSE	.33	TRUE SD	1.00	SEPARATION	3.06	Item RELIABILITY	.90
MODEL RMSE	.31	TRUE SD	1.01	SEPARATION	3.29	Item RELIABILITY	.92
S.E. OF Item MEAN = .28							

Figure 1 Pretest Problem Reliability



SUMMARY OF 30 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	16.0	15.0	-1.03	.45	.98	.0	.98	.0
P.SD	8.1	.0	1.56	.07	.32	.8	.32	.8
S.SD	8.3	.0	1.59	.08	.33	.9	.32	.8
MAX.	31.0	15.0	1.46	.63	1.74	1.7	1.74	1.6
MIN.	4.0	15.0	-3.75	.37	.54	-1.5	.42	-1.5

REAL RMSE	.48	TRUE SD	1.48	SEPARATION	3.07 Person RELIABILITY .90
MODEL RMSE	.46	TRUE SD	1.49	SEPARATION	3.26 Person RELIABILITY .91

S.E. OF Person MEAN = .29

Person RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .91 SEM = 2.42

SUMMARY OF 15 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	31.9	30.0	.00	.31	.97	-.2	.98	-.1
P.SD	11.7	.0	1.05	.03	.35	1.4	.32	1.2
S.SD	12.2	.0	1.09	.03	.37	1.5	.33	1.3
MAX.	58.0	30.0	1.80	.39	1.63	2.3	1.54	2.0
MIN.	15.0	30.0	-2.16	.27	.47	-2.5	.53	-2.2

REAL RMSE	.34	TRUE SD	.99	SEPARATION	2.96 Item RELIABILITY .90
MODEL RMSE	.31	TRUE SD	1.00	SEPARATION	3.19 Item RELIABILITY .91

S.E. OF Item MEAN = .28

Figure 2 Posttest Problem Reliability

The results of the reliability analysis of the Posttest questions were known that the Cronbach Alpha value was 0,91. This shows the reliability of the Posttest question in a very good category. The value of Reliability Measured Item is 0,90, it can be concluded that the quality of the items in the reliability aspect instrument is good. The value of Reliability Measured Person is 0,90, it can be concluded that the consistency of answers from students is of good reliability. The Separation Measured Item value of 3,06 can be concluded that the problem consists of 4 questions groups. The group of questions is a group of questions that are very easy, groups of questions are easy, groups of questions are difficult, and groups of questions are very difficult. Separation Measured Person value of 2,96 can be concluded that students consist of 4 groups of students. The group of students is a group of students with less ability, groups of students with low abilities, groups of students with sufficient abilities, and groups of students with good abilities. Mean Measured Person value of -1,03 shows that the average student ability tendency is smaller than the difficulty level of the problem [34].

The initial product revision phase, in the product revision phase based on limited trials, researchers did not make improvements. There are no suggestions for improvement from the limited trials that have been carried out. Field trial phase, in the field trial phase researchers, used two classes, namely class X MIA 1 and X MIA 2. Both classes were made into an experimental class. At the 1st meeting in MIA 1 class and MIA 2 class, 5 pretest questions were studied before learning and 5 posttest questions were tested after learning. Problem pretest and posttest tested consists of questions related to factual knowledge as much as 1 question is No. 1, related conceptual knowledge as much as 2 questions are No. 2 and No. 4, and related metacognitive knowledge as much as 2 questions are No. 3 and No. 5. The value of n-gain after learning the 1st meeting can be seen in Table 3.

Table 3 N-gain first meeting class X MIA 1 and X MIA 2

Aspect	Class X MIA 1			Class X MIA 2		
	Pre test	*Post test*	N-Gain	*Pre test*	*Post test*	N-Gain
Factual	2,171	4,200	0,530	2,114	2,971	0,221
Conceptual	1,686	2,971	0,298	1,457	2,443	0,217
Metacognitive	1,286	2,457	0,248	1,086	2,000	0,186

The meeting first in MIA 1 class obtained the N-gain value on factual knowledge of 0,530 showing an increase in HOTS factual knowledge of class X MIA 1 students in the medium category. The n-gain value on conceptual knowledge of 0,298 shows an increase in HOTS conceptual knowledge of class X MIA 1 students in low categories. The n-gain value of metacognitive knowledge is 0,248 which shows an increase in HOTS of metacognitive knowledge of class X MIA 1 students in low categories. This shows the need to improve the learning process in training HOTS ability of class X MIA 1 students on factual knowledge, conceptual knowledge, and metacognitive knowledge.

The meeting first in class MIA 2 n-gain values obtained on factual knowledge amounted to 0,221 show an increasing understanding of factual knowledge HOTS class X MIA 2 in the lower categories. The value of n-gain in the conceptual knowledge of 0,217 shows an increase in HOTS understanding of conceptual knowledge of class X MIA 2 students in low categories. The n-gain value on metacognitive knowledge is 0,186, showing an increase in HOTS understanding of metacognitive knowledge of class X MIA 2 students in low categories. This shows the need to improve the learning process in training HOTS ability of class X MIA 2 students on factual knowledge, conceptual knowledge, and metacognitive knowledge.

Problem at meeting first tested in class MIA 1 and class MIA 2 consists of questions related to factual knowledge as much as two questions are about No. 6 and question No. 8, a matter related to conceptual knowledge as much as 2 questions is a matter of No. 7 and No. 9, and about related metacognitive knowledge as much as 1 question that is a matter No. 10. The n-gain value after the execution of the learning process can be seen in Table 4.

Table 4.N-gain second meeting class X MIA 1 and X MIA 2

Aspect	Class X MIA 1			Class X MIA 2		
	Pre test	Post test	N-Gain	Pre test	Post test	N-Gain
factual	2,129	5,071	0,760	2,514	4,829	0,664
conceptual	1,914	5,043	0,766	2,271	4,857	0,566
metacognitive	1,429	3,657	0,488	1,429	3,657	0,488

The meeting second in class MIA 1 n-gain values obtained on the factual knowledge of 0,760 shows increased understanding of factual knowledge HOTS graders X MIA 1 in the high category. The value of n-gain in the conceptual knowledge of 0,766 shows an increase in HOTS understanding of conceptual knowledge of class X MIA 1 students in high categories. The n-gain value of metacognitive knowledge is 0,488 which shows an increase in HOTS understanding of metacognitive knowledge of class X MIA 1 students in the medium category. This shows the need for improvement of the learning process in training HOTS ability of class X MIA 1 students on students' knowledge of metacognitive knowledge [35].

The meeting second in class MIA 2 n-gain values obtained on the factual knowledge of 0,664 shows increased understanding of the factual knowledge HOTS class X MIA 2 in the categories of being. The value of n-gain in the conceptual knowledge of 0,566 shows an increase in HOTS understanding of conceptual knowledge of class X MIA 2 students in the medium category. The n-gain value on metacognitive knowledge is 0,186, showing an increase in HOTS understanding of metacognitive knowledge of class X MIA 2 students in the medium category. This shows the need for improvement of the learning process in training HOTS ability of class X MIA 2 students in the knowledge of metacognitive knowledge [36].

Problem at meeting second tested in class X MIA 1 consists of questions related to factual knowledge as much as 1 question is about No. 12, a matter related to conceptual knowledge as much as 2 questions is a matter of No. 13 and No. 14, and a matter related to the knowledge of metacognitive much as 2 questions namely problem number 11 and number 15. The n-gain value after the learning process can be seen in Table 5.

Table 5 N-gain third meeting class X MIA 1 and X MIA 2

Aspect	Class X MIA 1			Class X MIA 2		
	Pretest	Posttest	N-Gain	Pretest	Posttest	N-Gain
factual	4,171	5,314	0,750	3,771	4,686	0,410
Conceptual	2,686	4,357	0,504	2,814	3,943	0,354
Metacognitive	2,786	4,057	0,396	2,400	3,629	0,341

The third meeting in class MIA 1 n-gain value in factual knowledge of 0,750 shows an increase in HOTS understanding of factual knowledge of class X MIA 1 students on a high category. The value of n-gain in a conceptual knowledge of 0,504 shows an increase in HOTS understanding of conceptual knowledge of class X MIA 1 students in the medium category. The n-gain value of metacognitive knowledge is 0,396 which shows an increase in HOTS understanding of metacognitive knowledge of class X MIA 1 students in the medium category. This shows the need for improvement of the learning process in training HOTS ability of class X MIA 1 students on students conceptual and metacognitive knowledge.

The meeting third in class MIA 2 n-gain value on factual knowledge of 0,410 shows increased understanding of the factual knowledge HOTS class X MIA 2 in the categories of being. The value of n-gain in a conceptual knowledge of 0,354 shows an increase in HOTS understanding of conceptual knowledge of class X MIA 2 students in the medium category. The value of n-gain in metacognitive knowledge is 0,341 indicating an increase in HOTS understanding of metacognitive knowledge of class X MIA 2 students in the medium category. This shows an increase in HOTS ability in class X MIA 2 students in the medium category.

After the third meeting students were given HOTS questions to find out related to improving students high-level thinking skills. The questions given were 5 questions with 2 questions related to understanding the questions related to factual knowledge, namely Number 1 and Number 2, 2 questions related to conceptual knowledge, Number 3 and Number 4, and 1 question related to metacognitive knowledge, namely Number 15. After that, comparing the Pretest scores before students are trained to work on HOTS questions with the Posttest score after three student meetings is done with the HOTS question of the Polytomous Response model. The n-gain value after the learning process can be seen in Table 6.

Table 6 N-gain Instrument

Aspect	Pretest	Posttest	N-Gain	Average N-Gain
Factual	0,38	0,83	0,79	
Conceptual	0,22	0,82	0,77	**0,74**
Metacognitive	0,14	0,72	0,67	

The n-gain value in factual knowledge is 0,79, showing an increase in HOTS understanding of factual knowledge of class X students in high categories. The value of n-gain in a conceptual knowledge of 0,77 shows an increase in HOTS understanding of conceptual knowledge of class X students in high categories. The n-gain value of metacognitive knowledge is 0,67 which shows an increase in HOTS understanding of metacognitive knowledge of class X students in the medium category. The average n-gain value after students is trained using the assessment for learning instrument the polytomous response model is 0,74. This shows that there is an increase in HOTS ability of class X students using an assessment for learning instrument polytomous model response in the high category [37].

After the researchers conducted a field test, the researchers tested the practicality of the instrument for 3 high school teachers. Practicality test in this development, namely the review of assessment for learning instrument polytomous model response related to aspects of the model, substance, and use of instruments carried out by Examiners 1, Examiner 2, and Examiners 3. The results of the practicality test instruments were obtained using a practical questionnairequestionnaire. The data from the practicality test results can be seen in Table 7.

Table 7 Practical test results

Aspect	Tester Score			Average	Classification
	1	2	3		
Model	100,00%	75,00%	75,00%	83,33%	Very High
Substance	85,00%	80,00%	75,00%	80,00%	High
Use	87,50%	87,50%	75,00%	83,33%	Very High

The practicality test questionnaire contains 1 aspect of the instrument model. The average total score of the three testers regarding the fulfillment of the model aspects of the instruments that have been developed is 83,33% which means very high. The practicality test questionnaire contains 5 aspects regarding substance. The average total score of the three testers regarding the fulfillment of the substance aspect of the instrument that has been developed is 80,00% which means high. Practical test questionnaire contains 2 aspects regarding usage. The average total score of the three examiners regarding the fulfillment of the useful aspects of the instruments that have been developed is 83,33% which means very high. The value of each aspect which is more than 60%, means fulfilling the practicality criteria of the instrument [38].

Practical criteria are very high, there are aspects of the model and aspects of use. High practicality criteria are found in substance aspects. This, the instrument can be said to be practical and can be used in training students high-level thinking skills. Final Product Revision Phase, In the final product revision stage based on field tests, researchers did not make improvements. This is because there are no suggestions for improvement from the results of the field trials that have been conducted.

4. Conclusion

The conclusions in this study are an assessment for learning instruments the polytomous response model for practicing higher order thinking consists of the initial part of the front cover, foreword, table of contents. The middle part of the product development instrument consists of items, scoring rubrics. The final part of the development instrument consists of value recapitulation, literature list, back cover. Instrument development results have the contractual validity of 81,90% which is in the very high category, substance validity is 77,33% which is in the high category, and language validity is 84,45% which is in the very high category. The instrument of development results has a reliable value of 0,91 which is in the very good category. The instrument of development results also has a practicality aspect of the model of 83,33% which is in the very high category, the practicality of the substance is 80,00% which falls into the category very high, and the practicality value of the usage aspect is 83,33% which is in the very high category. Instrument development results have an n-gain value of 0,74, in this case, it can be concluded that the polytomous response model for assessment learning is effective in training students higher order thinking skills.

5. Acknowledgments

The author would like to thank the validator, which has been providing advice and improvements to the author.

References

[1] Mullis I V & Martin M O 2015 TIMSS 2015 Assessment Framework Chestnutt Hill *TIMSS &PIRLS International Study Center*

[2] Retnawati H 2011 Mengestimasi Kemampuan Peserta Tes Uraian Matematika dengan Pendekatan Teori Respons Butir dengan Penskoran Politomus dengan Generalized Partial Credit Model *Prosiding Seminar Nasional Penelitian Pendidikandan Penerapan MIPA Fakultas MIPA Universitas Negeri Yogyakarta*

[3] Kusuma M D 2016 Pengembangan Perangkat Instrumen Asesmen Higher Order Thinking Skill (HOTS) dalam Pembelajaran Fisika Siswa SMA pada Materi Fluida Statis *Universitas Lampung*

[4] Lewy, Zulkardi & Aisyah 2009 Pengembangan Soal untuk Mengukur Kemampuan Berpikir Tingkat Tinggi Pokok Bahasan Barisan dan Deret Bilangan di Kelas IX Akselerasi SMP Xaverius Maria

Palembang *JURNAL Pendidikan Matematika* **8** 2 15-28
[5] Heong Y M, Othman W D, Yunos J, Kiong T T, Hassan R, &Mohamad M M 2011 The Level of Marzano Higher Order Thinking Skills Among Technical Education Students International *Journal of Social and humanity* **1** 2 121-125
[6] Brookhart S M 2010 How to Assess Higher Order Thinking Skills in Your Classroom Alexandria *ASCD*
[7] Andrerson LW & Krathwohl DR 2001 A Taxonomy of Learning Teaching and Assessing *A Revision of Bloom's Taxonomy of Educational Objectives* New York Longman
[8] Limbach B & Waugh W 2010 Developing Higher Level Thinking*Journal of Instructonal Pedagogies* **2** 3 1-9
[9] Rosidin U 2016 *Penilaian Otentik* (Yogyakarta: Media Akademi)
[10]Widoyoko E P 2009 *Evaluasi Program Pembelajaran* (Yogyakarta: PustakaPelajar)
[11]Surapranata S 2005 *Analisis Validitas Reliabilitas dan Interprestasi Hasil Belajar* (Bandung: Remaja Rosdakarya)
[12]Haladyna T M 2004 Devoping and Validating Multiple Choise Test Items New Jersey *Lawrence Erlbaum Associates* **2** 1 22-32
[13]Brown S 2005 Assessment for Learning *Learning and Teaching in Higher Education* **1** 1 81-89
[14]Schuwirth L W T &Vleuten C P M V D 2011 Programmatic Assessment: From Assessment of Learning to Assessment for Learning *Medical Teacher* **2** 33 478-485
[15]Black P, Robert M C, Mary J & Pedder 2006 Learning How to Learn and Assessment for Learning: a Theoretical Inquiry *Research Papers in Education* **12** 1 39-54
[16]Stiggins R J 2010 Assessment Crisis: The Absence Of Assessment FOR Learning*A Special Section On Assessment* **7** 1 pp. 19-28
[17]Chappuis S &Stiggins R J 2002 Classroom Assessment for Learning *Educational Leadership* **60** 1 40-43
[18]Carless D 2005 Prospects for the implementation of assessment for learning *Assessment in Education* **12** 1 39-54
[19]Willian D, Clare l, Christine H, & Paul B 2004 Teachers Developing Assessment for Learning: Impact on Student Achievement *Assessment in Education* **11** 1pp. 49-65
[20] Black P, Robert M C, Mary J & Pedder D 2006 Learning How to Learn and Assessment for Learning: a Theoretical Inquiry *Research Papers in Education* **12** 1 39-54
[21] Harlen W 2005 Teachers' Summative Practices and Assessment for Learning-Tensions and Synergies *The Curriculum Journal* **16** 2 pp. 207-223
[22] Pedder D & Mary J 2012 Professional Learning as a Condition for Assessment for Learning *Assessment and Learning* **3** 2 33-48
[23] Isgiyanto A 2011 Diagnosis Kesalahan Siswa Berbasis Penskoran Politomus Model Partial Credit Pada Matematika *Jurnal Penelitiandan Evaluasi Pendidikan* **15** 2 309-325
[24] Susongko P 2010 Perbandingan Keefektifan Bentuk Tes Uraian dan Test let dengan Penerapan Graded Response Model (GRM) *Jurnal Penelitian dan Evaluasi Pendidikan* **14** 2 269-288
[25] Gall M D, Gall J P & Borg W R 2003 *Educational Research An Introduction (Seventh Ed)* Boston: Allyn and Bacon
[26] Sukmadinata N S 2011*MetodePenelitian Pendididkan* (Bandung: PT Remaja Rosdakarya)
[27] Sugiyono 2016 *Metode Penelitian Pendidikan* (Bandung: Alfabeta)
[28] Arikunto S 2016 *Prosedur Penelitian Suatu Pendekatan Praktik* (Jakarta: Rineka Cipta)
[29] Sudjana N 2012 *Penilaian Hasil Proses Belajar Mengajar* (Bandung: Remaja Rosdakarya)
[30] Meltzer D E 2002 The Relationship Between Mathematics Preparation and Conceptual Learning Gain in Phisics: A Posible Hidden Variable in Diagnostic Pretest Scores *American Journal Physics* **70** 2 1259-1267
[31] Ambarwati D & Suyatna A 2018 Interactive design for self-study and developing students' critical thinking skills in electromagnetic radiation topic *Journal of Physics: Conference Series* **948** 1 1-8
[32] 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* **1** 4 pp. 48-60
[33] Crenshaw P, Hale E, & Harper S L 2011 Producing intellectual labor in the classroom: The utilization of a critical thinking model to help students take command of their thinking *Journal of College Teaching and Learning* **8** 7 13-26
[34] Snyder L G & Snyder M J 2008 Teaching critical thinking and problem solving skills *The Journal of Research in Business Education* **50** 290
[35] Peter E E 2012 Critical thinking: Essence for teaching mathematics and mathematics problem solving skills *African Journal of Mathematics and Computer Science Research* **5** 3 39-43
[36] Hyytinen H, Nissinen K, Ursin J, Toom A, & Lindblom-Ylänne S 2015 Problematising the equivalence of the test results of performance-based critical thinking tests for undergraduate students *Studies in Educational Evaluation* **44** 1-8
[37] Fakhriyah F 2014 Penerapan Problem Based Learning dalam Upaya Mengembangkan Kemampuan Berpikir Kritis Mahasiswa *Jurnal Pendidikan IPA Indonesia* **3** 1 95-101
[38] Dixon F A, Prater K A, Vine H M, Wark M J, Williams T, Hanchon T & Shobe C 2004 Teaching to their thinking: A strategy to meet the critical-thinking needs of gifted students *Journal for the Education of the Gifted* **28** 1 56-76
[39] Agustini, R & Suyatna, A 2018 Developing Inquiry-Based Practice Equipment of Heat Conductivity to Foster the Students' Critical Thinking Ability *Jurnal Ilmiah Pendidikan Fisika Al-BiRuNi*, **7** 149–57