Research Article doi: 10.12973/eu-jer.8.3.743

European Journal of Educational Research

Volume 8, Issue 3, 743 - 751.

ISSN: 2165-8714
http://www.eu-jer.com/

The Development of an Instrument to Measure the Higher Order Thinking Skill in Physics

Syahrul Ramadhan*
Yogyakarta State University,
INDONESIA

Djemari Mardapi
Yogyakarta State University,
INDONESIA

Zuhdan Kun Prasetyo
Yogyakarta State University,
INDONESIA

Heru Budi Utomo
Yogyakarta State University,
INDONESIA

Received: March 27, 2019 ▪ Revised: May 20, 2019 ▪ Accepted: June 13, 2019

Abstract: This study is conducted to develop the diagnostic test, which can be used to measure the higher-order thinking skill (HOTs) of students of first-grade senior high school in Bima district, West Nusa Tenggara. The step of developing instruments such as test which using modification model of Oreondo which include two activities such as test designing and test trials. The analysing technique of validity of content used Aiken formula, classical test theory used software Iteman 4.3, the model of Rasch used software Winstep and analysing reliability used software SPSS. The conclusion which can be taken are developing instrument has the characteristics as a useful instrument and fulfil requirement used to measure. This case proved from the data of analysis result which confirm that the instrument has been achieved the content of validity by expert judgment and obtained the empirical evidence, both as classical test theory or Rasch model.

Keywords: *Higher order thinking skill, physics, instrument.*

To cite this article: Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher order thinking skill in physics. *European Journal of Educational Research*, *8*(3), 743-751. https://doi.org/10.12973/eu-jer.8.3.743

Introduction

Improving the students' Higher Order Thinking Skill (HOTS) in physics subject can be done through giving appropriate assessment because by using appropriate assessment can be encouraged students to learn by Higher Order Thinking in Bloom's taxonomy which has been revised by Anderson et al. (2001). An ability which includes into Lower Thinking Order (LOT) is the ability to remembering understanding and applying while Higher Order Thinking (HOT) including analysing, evaluating, and creating. Therefore, an appropriate assessment is not only measuring the low thinking order but also the Higher Order Thinking, which includes students' physics ability in analysing, evaluating, and creating.

Generally, there still many teachers who aren't familiar with the test based the Higher Order Thinking. Whereas, evaluating by Higher Order Thinking should have been begun to be introduced in assessing the process by the teacher in the classroom. Therefore, the ability in designing and developing the test based on the Higher Order Thinking should be owned by the teachers. If the teachers; skill is low in creating a test based the Higher Order Thinking skill, it will effect on the low quality of test produced, and it will be an adverse effect on the process of measuring and evaluating the competency of students. The statement above is same with the argument from Ong, Hart, and Chen (2016) which state that how essential of teachers'role in helping students to build their scientific ideas and their reflective thinking skill.

Developing the test level of Higher Order Thinking is not easy. The resulting study from Jensen, McDaniel, Woodard, and Kummer (2014) showed that in writing the test level of HOT is a challenging task for teachers, and it needs to be improved because it really will help students in obtaining the deep understanding toward the materials thought. Therefore, from those descriptions, the researcher considers that it needs to make a set of test which can be used by physics teacher in both of evaluating the process and as guiding in forming the test level of HOTs.

* **Corresponding author:**
Syahrul Ramadhan, Yogyakarta State University. Jalan Colombo No. 1, Karangmalang, Yogyakarta 55281, Indonesia. ✉ syahrul.ramadhan2015@student.uny.ac.id, laarul892@gmail.com

Definition of Higher Order Thinking Skill (HOTS)

Higher Order Thinking has been learnt in a long period, and it is not a new topic recently appeared (Wang & Wang, 2011, 2014), there are many definitions, level and level developed relate to Higher Order Thinking. Learning by involving Higher Order Thinking skill is believed can improve students' ability in preparing themselves facing the challenging and expanding era, as well students more capability to live in social life (Chetty, 2015; Snyder & Snyder, 2008; Ten Dam & Volman, 2004), therefore learning which also involving Higher Order Thinking skill really need to be developed.

Higher Order Thinking Skills is defined including critical thinking, logic, reflective, metacognitive, and creative (Limbach & Waugh, 2010; Wang & Wang, 2014). All those skills will be active when someone faces an unusual problem, uncertainty, question and choice. The successful applying from these skills contained in explanation, decision, appearance and a valid product. The application is inappropriate with the context from the knowledge and experience as well as advanced developing or another intellectual ability.

Snyder and Snyder (2008) give steps in improving students' thinking skill by 1) using an active learning strategy which involves students in the learning process rather than relies on lecturing and remembering 2) instruction focus on the learning process rather than on content and 3) using assessment technique which gives students intellectual challenge rather than memory.

Higher order thinking occur when someone takes new information and information saved in memory and related each other and expands this information to reach the goal or finding the possible answer in a confusing situation. Higher Order Thinking can achieve various goals. Deciding what should be believed; determining what should be done; creating a new idea, a further object, or artistic expression; making a prediction, and solving a problem.

Giving learning to improve Higher Order Thinking skill is not an easy way. There are several obstacles which need to be anticipated by teacher, teaching is considered as knowledge transmission, learning HOTs involving coverage specific content but expanding, teacher's expectation toward students is low, big classroom, the lack of teacher planning time teaching culture which isolate teacher (Fischer, Bol, & Pribesh, 2011). Other several things such as gender, academic prestige and social, economic status don't affect significantly toward the enhancement Higher Order Thinking skill for students (Ozsoy-Gunes, Gunes, Derelioglu, & Kirbaslar, 2015).

In Indonesia, Higher Order Thinking skill is always linked with Bloom's taxonomy revised, mainly top three such as C4 (analysing), C5 (evaluating), C6 (creating). Even the use of Bloom's taxonomy is also loaded in the curriculum used in Indonesia. Therefore, the concept of Higher Order Thinking skill used in this study refers to the idea of higher-order thinking from Bloom's taxonomy revised.

Research purposes

In this study, physics materials used for the instrument developed is vertical irregular motion. The reason why this material used is because of the researcher experience as a teacher, straight irregular motion material is just considered as rote of formula. The teacher cannot make creation by presenting the example of a question and contextual stimulus, therefore, the researcher needs to develop this instrument which can be used by the teacher in learning. So, the purposes of this research are;

- To find out the validity of the content (quantitative or qualitative) of HOTs physics Instrument which is developed.
- To find out empirical validity (item difficulty, item discrimination and distractor) of HOTs physics Instrument which is developed.
- To find out the reliability of HOTs physics Instrument which is developed,
- To find out the empirical validity (using the Rasch Model) of HOTs physics Instrument which is developed.

Method

Type of Research

This study is conducted to develop the diagnostic test which can be used to measure the higher order thinking skill, as well noticing the higher order thinking skill of students of first-grade senior high school N in Bima district, West Nusa Tenggara. The step is done to begin with developing an instrument, analysing the characteristic of the instrument and looking at the description of HOTs of students' ability. The step of developing instruments such as test which using modification model of Oreondo (1984) which include two activities such as test designing and test trials.

Designing test consist of several activities such as determination the aim of test, determination of the competency which will addressed, determination of material tested, preparation the lattices, test writing based on the guiding of developing test model of higher order thinking skill, validity of test item, improvement item and assembly test as well preparation score guidelines. The step for the trial test is done after the test formed.

Test trial is conducted in two schools such as first-grade students of SMAN 1 Sape and first-grade students of SMAN 3 Sape Bima District West Nusa Tenggara, with the total of sampling used is 201 numbers of students.

Table 1. The step of developing instruments

Developing instruments	Step/activities
Test Designing	Determination the aim of test Determination of the competency which will addressed Determination of material tested Preparation the lattices Test writing Validity of test item Improvement item Assembly test Preparation score guidelines
Test Trials	Test trial is conducted in two schools Total of sampling used is 201 numbers of students

Data Analysis

Analysing technique of validity of content used Aiken (1985) formula by excel program, for empiric validity, classical test theory used software ITEMAN 4.3, while for analysing empiric validity model of Rasch used software winstep. Analysing reliability used Cronbach Alpha used software SPSS and to portrait students ability in using statistic descriptive by using excel.

Content of Validity

This validity is determined by using the experts' deal. The expert of the field of study is deal or domain which is measured to determine the level of validity of content (Heri Retnowati,2016). This case is caused by instrument measurement such as a test or questionnaire which is valid proofed if the expert believes that the instrument could measure the mastering of skill which is defined in domain measured. This analysing validity used Aiken formula.

Aiken formulates the Aiken's formula V to count content-validity coefficient based on assessment result from the expert panel as much as n people toward an item from the terms of how far the item represents the measured contract. The submitted formula by Aiken can be shown below (Azwar,2012):

$$\boxed{V = \frac{\Sigma s}{n(c-1)}}$$

Where,

V = validity index item

S = score applied, each rater reduced low score in category used ($s=r-l_o$, → r = rater score choice and l_o = low score in score categorizing)

N = number of rater

C = number of criterion/rating

Reliability

The instrument's reliability is intended to see the consistency of the tests made if the observation is repeated. The level of instrument's reliability empirically proven by the amount of the reliability coefficient which is in the range of 0 to 1 (Mardapi, 2008, 2012). The higher the coefficient value means the higher the reliability, and vice versa. The coefficient formula of *Alpha Cronbach*'s used to estimate the test reliability and calculate it using *Iteman 4.3* computer program. The reliability estimation is based on the index of instrument reliability that is good if > 0,7 (Mardapi, 2008).

Classical Test Theory

Classical test theory used a simple mathematic model to show the relation among observation score, the real score and error score. This model is followed with the number of assumptions to simplify the formula in estimating reliability index and the validity of the instrument. Classical test theory develops the measuring model to assess the magnitude of parameter skill and the parameter of the item. The parameter of ability is declared as the number of the right items for the form of multiple choices or the number of the score for the form of the briefing. The parameter of items is the item difficulty, item discrimination, and distractor.

Quantitative analysis of the characteristics of the HOTs physics Instrument was conducted based on the classical Test Theory approach. The researcher analyses the students' response patterns (based on student answer sheets) to see the information about test instruments that are proper and not proper to be tested based on the item parameters, that are the Item difficulty, Item discrimination, and the distractor.

The analysis of the Item difficulty of HOTs physics Instrument used *Iteman 4.3* computer program. The Item difficulty of the question can be seen in the column *Prop. Correct.* Item questions that have Item difficulty are in the interval of 0.3 to 0,8.

The analysis of Item discrimination of HOTs physics Instrument can be seen in the *Point Biserial* column conducted using the *Iteman 4.3* computer program. The criteria for good questions have a value of $D \geq 0.3$, while the question that has a value of $D \leq 0.3$ need to be revised or replaced with new items.

The information on the distractors can also be taken from *Iteman 4.3* computer programs, that is in the column *Prop Endorsing*. The distractors are said to function if the value of *PropEndorsing* in each multiple choice has a higher value than 0.05. The proportion value from each question that has a value less than the value of *PropEndorsing*, the distractor needs to be revised.

Rasch Model

Rasch model provides a framework which used to be explained to a researcher in human knowledge which will compare their data. The reader will not surprise to know that the formulation of measuring and more detail model and more suitable which described in an analogic way.

The opportunity (conventional is written as P) from the successful (the result from 1 and 0) is one different function between the ability (we used B) and the Item difficulty (we symbolise D); or

$$\boxed{P_{ni}(x=1) = f(B_n - D_i)}$$

The logistic curve which introduced is known as expectation response curve from the Rasch model for separation item (it called items characteristics curve or ICC). By logits (log possibility unit), and centimetres in X-axis, summarize Bn-Di, the difference of each couple of people- items from analogic path, we can read the estimate of successful possibility if there is a person and item meet the distribution of differential item which guided on scale, by only knowing Bn-Di the differences between the ability and the difficulty. Theorem measurement for Rasch model which need matrix data from the test and so on meets the expectation of possibility from two formal statements of Rasch model; first as formula (1) above, and second in the form of graphic.

This fit is a principle of quality control which used to help to decide whether the person's ability or item is near enough with the necessities of Rasch model to be considered as scale step of interval linear. The differences between the person, between item, person and item can be read directly from the level of interval scale to make a comparison which interpreted as "how many differences" between two locations probability condition.

Results

Analysis Result of the Content of Validity

"To find out the validity of the content (quantitative or qualitative) of HOTs physics Instrument which is developed."

Three experts conduct expert judgment in this study in the physics field, research and academic evaluation. An instrument can be said valid if the expert believes that the instrument measures the things that will be measured. The expert judgment gives the assessment which will be used to prove the content of validity. Providing values for each item uses numbers 1, 2, 3 and 4. Number one represents invalid, two represents less valid, three represents valid, and four represents very valid. The result showed as can be seen in the table below:

Table 2. The analysis result of Aiken

Items	Rater 1	Rater 2	Rater 3	Value V
1	4	4	4	1
2	3	3	3	0,67
3	4	4	4	1
4	4	4	4	1
5	4	4	4	1
6	3	3	3	0,67
7	4	4	4	1
8	3	3	4	0,78
9	4	4	4	1
10	4	4	4	1
11	4	4	4	1
12	4	4	4	1
13	4	4	4	1
14	4	4	4	1
15	4	4	4	1
16	4	4	4	1
17	4	4	4	1
18	4	4	4	1
19	4	4	4	1
20	4	4	4	1
21	4	4	4	1
22	4	4	4	1
23	4	4	4	1
24	3	3	4	0,78
25	4	3	3	0,78

Based on the data above, the range of value for the instrument is 0,67-1. While based on Aiken's table, if the number of items is 25 and three raters, so the minimum limit that can be accepted is 0,66. Based on the data it can be said that all of the items proved to be validly reviewed from the content of validity.

Table 3. Reliability with Formula Intraclass Correlation Coefficient

Cronbach's Alpha	N of Raters
0,881	3

The formula of the Intraclass Correlation Coefficient is conducted to estimate the reliability of instrument; analysis is done by using SPSS software. The result is 0,881 and more significant than 0,70, therefore, it can be said that based on the analysis of data of expert's deal the instrument developed is reliable.

Analysing by Classical Test Theory

> *"To find out the reliability of HOTs physics Instrument which is developed,"*
>
> *"To find out empirical validity (item difficulty, item discrimination and distractor) of HOTs physics Instrument which is developed."*

Analysing the item of HOTs test sub-section straight irregular motion for the first-grade student of SMAN I Sape in first-semester academic year 2016/2017 is conducted with support by statistic software which is ITEMAN 4.3.0.3 version. From the analysis result can be obtained the characteristic of question items. Look at the analysis below.

T*able 4. Reliability*

Score	Alpha
Scored items	0,810

From the analysis result by using ITEMAN, it is obtained that the value of reliability is about 0.883. This value is big enough and indicates that the instrument is reliable. Therefore it can be used to be a measuring instrument. Besides reliability, there are several aspects or items characteristics which are needed to be noticed in classical test theory such as the Item difficulty, Item discrimination and distractor.

The Item difficulty is defined as the proportion of test participant that response the right answer in certain test item and its value about 0 to 1 (Mardapi, 2012; Ramadhan & Mardapi, 2015). The analysis result for the Item difficulty is:

Table 5. Frequency of Distribution for the Item difficulty

Score	Frequency
0,0 to 0,1	0
0,1 to 0,2	0
0,2 to 0,3	0
0,3 to 0,4	0
0,4 to 0,5	6
0,5 to 0,6	17
0,6 to 0,7	2
0,7 to 0,8	0
0,8 to 0,9	0
0,9 to 1,0	0

The instrument made in a test form is on the average level for the Item difficulty. Therefore the instrument categorises well. The Item difficulty of the items can be accepted if the magnitude is 0,30 to 080 (Mardapi, 2012; Ramadhan & Mardapi, 2015).

The index of Item discrimination is defined as the deviation between the proportion of the right answer in high group and the proportion of the correct answer in low group (Mardapi, 2008). Classifying these groups can be done by various method depends on its need, the Item discrimination of the items can be accepted if the magnitude is less than 0, 30 (Mardapi, 2012; Ramadhan & Mardapi, 2015). The analysis of Item discrimination to this research can be seen in the *Point Biserial* column conducted using the *Iteman 4.3* computer program. The criteria for good questions have a value of D≥0.3, while the question that has a value of D≤0.3 need to be revised or replaced with new items. The analysis result of Item discrimination can be seen below.

Table 6. The frequency of distribution for Item discrimination

Score	Frequency
0,0 to 0,1	0
0,1 to 0,2	1
0,2 to 0,3	8
0,3 to 0,4	10
0,4 to 0,5	2
0,5 to 0,6	4
0,6 to 0,7	0
0,7 to 0,8	0
0,8 to 0,9	0
0,9 to 1,0	0

Item discrimination can be accepted if the magnitude is less than 0,30 (Azwar, 1997, 2007). Based on the statement and the trial test result, there are nine questions which have Item discrimination less than 0,3. While sixteen questions have good Item discrimination. Based on reviewed toward the Item difficulty, Item discrimination, and Distractor, so it can be concluded that the analysis result of items can be summarised on the table below:

Table 7. Classical Analysis Result

No	Decision	Item	N
1	**Received**	1, 3, 4, 5, 8, 10, 12, 14, 15, 17, 18, 19, 21, 22, 23, 24	16
2	**Received with revision**	2, 6, 7, 9, 11, 13, 16, 20, 25	9
3	**Deleted**	-	0

Based on the empirical test, it is obtained that there are nine questions accepted with the requirement and sixteen items of questions received. The reason it is received with conditions is caused by Item discrimination which less than 0,3 as well as functioning from destructor where the option is chosen by less than 5% of the sample. The question accepted by this requirement then revised again to produce question which has quality.

Analysing by Rasch Model

"To find out the empirical validity (using the Rasch Model) of HOTs physics Instrument which is developed."

Examination of proper determination in each item toward the model follows the rule of Adams and Kho (1996), an item of question fit with the model if the value of INFIT MNSQ is between 0,77 to 1,30. Analysis result with winstep program presented on the table below:

Table 8. Analysis Result of Rasch Model

	Total	Count	Measure	Model	In fit		Outfit	
	Score			Error	Mnsq	Zstd	Mnsq	Zstd
MEAN	13,1	25	0,13	0,46	1	0	1	0
S.D.	4,8	0	0,98	0,11	0,08	1	0,12	1
MAX.	24	25	3,22	1,02	1,17	2,9	1,31	2,9
MIN.	1	25	-3,21	0,4	0,83	-2,7	0,6	-2,7

The value of reliability is 0,79 accepted standard is 0,65. The validity empirically proved by the goodness of fit toward the Partial Credit Model (PCM). Based on the table, it is obtained the information of the average value, and standard deviation of INFIT MNSQ are same which each 1,00and 0,08 (almost 0,00), so the fittest with PCM is 1 PL. This case means that the test empirically valid. Another proof is supported by all of the items which have INFIT MNSQ value which is 0,83 to 1,17 which is between the value limit acceptance using INFIT MNSQ or fit according to model (between 0,77 to 1,30) it means all of the items which are 25 items is fit.

Discussion

This study is conducted to develop a diagnostic test which can be used to measure the higher order thinking skill, as well to take a portrait students' higher order thinking the skill of first-grade students of SMA N in Bima District West Nusa Tenggara. In Indonesia, higher order thinking skill is always linked with Bloom's taxonomy revised, mainly on the top three such as C4 (analysing), C5 (evaluating) and C6 (creating). Even the used of Bloom's taxonomy loaded in the curriculum used in Indonesia. Therefore, the concept of HOTs used in this study refers to the concept of higher-order thinking from Bloom's taxonomy revised.

In this study, physics materials used for developing instrument is straight irregular motion. The reason of researcher used this material because of the experience of the researcher as a teacher and this material assumed that it just remembers the formula. The teacher cannot be creative by presenting the sample of a question and contextual stimulus. Therefore the researcher needs to consider that developing this instrument can be used by the teacher in learning.

The verification of validity and reliability from this instrument is by the validity of content, classical test theory, and Rasch model. The content of validity used expert judgment which conducted by three experts in the physics field, academic research and evaluation. The result of the expert's validity showed that the susceptible Aiken's value for the instrument is 0,67-1. While based on Aiken's table, if the number of items is 25 and raters is three, so the minimum limit received is 0,66. Based on the data it can be said that all of the items are validly reviewed from the content of validity. Polit and Beck (2006) argue that the clarity about content validation in the study of instrument development was essential. The reliability from the expert is then analysed by the formula of Interclass Correlation Coefficient by SPSS. The result is 0,881 and more significant than 0,7. Therefore it can be said that based on analysing the data of the expert deal that developing an instrument is reliable (Mardapi, 2012).

The next verification is using classical test theory by considering the Item difficulty, Item discrimination and the distractor. The Item difficulty of the items can be approved if the magnitude is about 0,30 until 0,80, while Item discrimination can be accepted if the scale is less than 0,30 and the distribution of answer minimum 5% in distractor (Mardapi, 2012). Based on the empirical result test by classical test theory it is obtained the information that there are nine answers received by requirement and sixteen items of questions are received. The reason received by the condition is caused by the value of Item discrimination which is less than 0.3 as well as the functioning of distraction where the option is chosen by less than 5% of the sample. The question received with this requirement, then revised again to produce the question which has a quality, but as a whole, there isn't bad question or discard. The analysis result by using items obtained the value of reliability which is about 0,883. This value is big enough and indicates that the instrument is reliable.

Empirically validity used the Rasch model proved by the goodness of fit toward the Partial Credit Model (PCM). Based on the table obtained the information of average value and standard deviation INFIT MNSQ is same which each 1,00 and 0,08 (almost 0,00) so the test fit by PCM 1 PL. This case means that the test empirically valid. The other proof is supported by all of the items which have the value of INFIT MNSQ from 0,83 to 0,79 which lies between item acceptance items using INFIT MNSQ or fit according to model (between 0,77 to 1,30) it means that all of the items are about 25 items of all fit. The value of validity of instrument based on Rasch analysis model with win step is 0,79, the standard of acceptance is 0,65. Therefore the instrument can be used as a measuring instrument. This instrument has high strength because it is composed of items that have high information functions (Hambleton, Swaminathan, & Rogers, 1985). This can occur because this test is in accordance with the ability of students who are tested (Istiyono, 2013).

Based on the description above, the instrument has conducted by the content of validity through expert judgment and got the proof of validity empirically. Besides that, the reliability of the test is high. More than that this instrument proved can be used to measure the higher-order thinking the skill of physics on the material of straight irregular motion. Therefore, this instrument has been fulfilled; the requirement is used for measuring the higher order thinking level of physics of the first-grade student on the straight irregular motion material. Tests that contain valid HOTS-level questions encourage students to think about the subject matter (Barnett & Francis, 2012; Istiyono, Mardapi, & Suparno, 2014).

Conclusion

Based on the description above, the conclusion which can be taken in developing instrument has the characteristics as a useful instrument and fulfil requirement used to measure. This case proved from the data of analysis result which confirm that the instrument has been achieved the content of validity by expert judgment and obtained the empirical evidence, both as classical test theory or Rasch model.

It cannot be denied that there are many weaknesses from this study which can be a change to develop later by the researcher or others, such as the scope of material used is still narrow. Many mistakes done by students because of the researcher make mistakes on the process of counting convert which is caused by a wrong concept. Moreover, students tend to assume that the form of the question is still relatively new for them, that means that assessment using a test level of HOTs still rarely used by the teachers. Besides that, the researcher needs to consider that the study needs to reach the learning model which can be improved students' higher order thinking skill.

Acknowledgements

This work was financially supported by Lembaga Pengelola Dana Pendidikan (LPDP). We would like to express our institution, Yogyakarta State University that has provided many experiences in term of research. Our deepest gratitude to students who participated in our experiment research and also who have contributed to the collection of our empirical data. Thankfully to the teacher of senior high school who helped this research

Conflict of Interests

The authors declare no conflict of interest.

References

Adams, R., & Kho, S.-T. (1996). *Acer quest version 2.1*. Camberwell, Victoria: The Australian Council for Educational Research.

Aiken, L. R. (1985). Three coefficients for analysing the reliability and validity of ratings. *Educational and psychological measurement, 45*(1), 131-142.

Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., . . . Wittrock, M. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy*. New York. Longman Publishing.

Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. *Cognition and Instruction*, 9(2), 137-175.

Azwar, S. (1997). Validitas dan reliabilitas *(Validity and reliability)*. Yogyakarta, Indonesia: Pustaka Pelajar.

Azwar, S. (2007). Validitas dan reliabilitas *(Validity and reliability)*. Yogyakarta, Indonesia: Pustaka Pelajar.

Barnett, J. E., & Francis, A. L. (2012). Using higher order thinking questions to foster critical thinking: A classroom study. *Educational Psychology, 32*(2), 201-211.

Chetty, N. (2015). Teaching Teachers to Teach Physics to High School Learners. *Procedia-Social and Behavioral Sciences, 174*, 1886-1899.

Fischer, C., Bol, L., & Pribesh, S. (2011). An investigation of higher-order thinking skills in smaller learning community social studies classrooms. *American Secondary Education, 39*(2), 5-26.

Hambleton, R., Swaminathan, H., & Rogers, H. (1985). *Principles and applications of item response theory.* Boston, MA: Kluwer-Nijhoff Publishing Company.

Istiyono, E. (2013). Tes Kemampuan Berpikir Tingkat Tinggi Fisika di SMA Langkah Pengembangan dan Karakteristiknya (Test of Physics High Order Thinking Skill at Senior High School: Developmental Steps and Its Characteristics) (Doctoral dissertation, Universitas Negeri Yogyakarta), Retrieved from https://eprints.uny.ac.id/22731/1/Artikel%20PDD%20PhysTHOTS.pdf.

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi fisika (pysthots) peserta didik SMA (Developing Higher Order Thinking Skill Test Of Physics (Physthots) For Senior High School Students) . *Jurnal Penelitian dan Evaluasi Pendidikan, 18*(1), 1-12.

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test… or testing to teach: exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*(2), 307-329.

Limbach, B., & Waugh, W. (2010). Developing higher level thinking. *Journal of Instructional Pedagogies, 3*, 1.

Mardapi, D. (2008). *Teknik penyusunan instrumen tes dan nontes (Techniques for preparing instruments test and non-test)*. Jogjakarta, Indonesia: Mitra Cendekia.

Mardapi, D. (2012). *Pengukuran penilaian dan evaluasi pendidikan (Education measurements, assessment and evaluation)*. Yogyakarta, Indonesia: Nuha Medika.

Ong, K. K. A., Hart, C. E., & Chen, P. K. (2016). Promoting higher-order thinking through teacher questioning: a case study of a singapore science classroom. *New Waves, 19*(1), 1-19.

Oreondo, L. L. (1984). *Evaluating educational outcomes*: *Manila, Philippines:* Rex Bookstore, Inc.

Ozsoy-Gunes, Z., Gunes, I., Derelioglu, Y., & Kirbaslar, F. G. (2015). The reflection of critical thinking dispositions on operational chemistry and physics problems solving of engineering faculty students. *Procedia-Social and Behavioral Sciences, 174*, 448-456.

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health, 29*(5), 489-497.

Ramadhan, S., & Mardapi, D. (2015). Estimasi Kesalahan Baku Pengukuran Soal-Soal UAS Fisika Kelas XII SMA di Kabupaten Bima NTB (the estimation the standard error of measurement in physics end-of-semester tests of senior high schools in kabupaten Bima, NTB). *Jurnal Evaluasi Pendidikan, 3*(1), 90-98.

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir (Validity, reliability and item charactheristic).* Yogyakarta, Indonesia: Parama Publishing.

Snyder, L. G., & Snyder, M. J. (2008). Teaching critical thinking and problem solving skills. *The Journal of Research in Business Education, 50*(2), 90.

Ten Dam, G., & Volman, M. (2004). Critical thinking as a citizenship competence: teaching strategies. *Learning and instruction, 14*(4), 359-379.

Wang, S., & Wang, H. (2011). Teaching higher order thinking in the introductory MIS course: A model-directed approach. *Journal of Education for Business, 86*(4), 208-213.

Wang, S., & Wang, H. (2014). Teaching and learning higher-order thinking. *International Journal of Arts & Sciences, 7*(2), 179.