The 4th International Conference on Technology, Education, and Sciences
The Institute of Research and Community Service, Universitas Sarjanawiyata Tamansiswa, Yogyakarta, Indonesia

InCoTES 8 November 2022 ISSN: XXXX-XXXX

Analysis of HOTS-Based Assessment Instruments for Science Subjects

Evi Hikmah Nurchayati[1], Yuli Prihatni [2], Rahayu Retnaningsih [3]
Universitas Sarjanawiyata Tamansiswa, Indonesia [1-3]

evihikmah2018@gmail.com

Keywords

HOTS, Research, and Development

Abstract

This study examines the Higher Order Thinking Skill (HOTS) Instrument for Science Subjects in Class IX. This activity aims to identify the appropriate and standard HOTS instrument for science subjects for grade IX students to assess higher-order thinking skills. The research and development methodology is used in this study. In this study, the HOTS instrument went through the following stages of research and development: (1) research and information gathering, (2) planning, (3) initial product development, (4) limited product trial, (5) product revision, (6) field trials, and (7) final product revision. The qualitative and quantitative data analysis was used to determine the HOTS instrument's validity, reliability, discriminatory index, difficulty index, and distractor quality criteria. The development's final result demonstrated that the HOTS instrument, in the form of multiple-choice questions with 24 items, is usable. The HOTS instrument was declared valid based on expert evaluation, with Aiken validity results greater than 0.8 in the outstanding category. The HOTS instrument has a difficulty level dominated by 50% of moderate questions, 70% of item discriminating power including having adequate discriminating power, and the effectiveness of the distractor; as much as 47.5% of the questions were in the sound and excellent categories, with a reliability of 0.85 in the very high category.

Introduction

Permendikbud No. 21, 2016, concerning the Graduate Competency Standards, states that the reference for the standards is Bloom Taxonomy which was first introduced by a group of researchers led by Benjamin Bloom in 1956 and further developed by Anderson and Krathwol in 2001. Bloom Taxonomy categorizes achievement learning is divided into three domains, namely the knowledge dimension, the attitude dimension related to the mastery of attitudes and behavior, and the skill dimension related to the mastery of skills. Dimensions of knowledge are classified into factual, conceptual, procedural, and metacognitive. This cognitive dimension is arranged hierarchically, starting from remembering, understanding (understanding), applying (applying), and analyzing (analyzing). Evaluating and creating (Wiwik Setiawati et al.,2019: 15)

To measure the learning achievement of the cognitive dimension using an assessment or assessment, namely making decisions regarding information obtained from the learning process and other supporting aspects of educational activities (Uno & Koni, 2012: 2 in Siti Fatimah, 2020: 320). Meanwhile, Sunarti & Rahmawati (2014:7) explain the notion of assessment as a systematic and continuous process to obtain information about learning and learning outcomes. Arifin (2019: 4) explains the meaning of assessment, which is a collection of information about the results and data collected to make a decision on the learning process.

In conducting the assessment using an instrument. Understanding Instruments, according to Arikunto (2010: 203), states that "instruments are tools that are selected and used by researchers in their activities to collect data so that these activities become systematic and facilitated by them." An evaluation tool or instrument in Arikunto (2012: 40-51) is something that can be used to make it easier for someone to carry out tasks or achieve goals more effectively and efficiently. Sudjiono (2011: 4) explains "judging is the activity of making decisions on something by basing oneself or holding on to good or bad, healthy or sick, smart or stupid, and so on."

Based on the opinions of some of these experts, it can be said that an instrument is a tool used to collect information about the variables being studied. In this case, the instrument in question is an assessment instrument. Assessment is a systematic process involving gathering information (numbers or verbal descriptions), analysis, and interpretation to make decisions. Therefore, based on the understanding of the instrument and assessment, it can be concluded that an assessment instrument is a tool used in collecting data that is used as a basis for analysis and interpretation for decision-making.

In the cognitive domain, the 2013 curriculum on content standards is designed so that students have the ability to think critically, creatively, logically, and analytically in order to be able to compete internationally. In addition, Higher Order Thinking Skills (HOTS) are one of the abilities in the cognitive domain that is currently a concern in the 2013 curriculum. The standard of assessment is emphasized learning outcomes that focus more on higher-order thinking skills (Kemendikbud, 2017). According to Ernawati (2017:196-197), Higher Order Thinking Skills (HOTS) is a way of thinking that no longer only memorizes verbally but also interprets the nature contained among them; to be able to interpret meaning requires an intergalactic way of thinking. By analyzing, synthesizing, and associating to draw conclusions toward creating creative and productive ideas. High Order Thinking Skills (HOTS) is the ability to connect, manipulate and transform existing knowledge and experience to think critically and creatively in order to solve problems in new situations (Rofiah et al., 2013 in Fitriana, 2020: 886). Higher Order Thinking Skills (HOTS) include two main characteristics, namely the ability to think critically and think creatively (Conklin, 2012: 14). The characteristics of HOTS revealed by Resnick (in Budiman & Jailani, 2014: 141) include non-algorithmic, complex nature, multiple solutions (many solutions), involving variations in decision making and interpretation, application of multiple criteria (many criteria), and being effortful requires much effort).

Meanwhile, according to Brookhart (2010: 14) higher order thinking skills (HOTS) include the ability to analyze, evaluate and create, logical reasoning (logical reasoning), decision making (judgment), critical thinking, problem solving, creativity and creative thinking. HOTS can be said as learning skills to communicate, reasoning skills, problem solving and learning systematically by connecting existing ideas, and connecting positive attitudes towards a goal.

(Yaniawati, 2013:110). Bloom's taxonomy in the cognitive domain is the basis for higher order thinking skills or known as Higher Order Thinking Skills (HOTS). Dimensions of cognitive processes in Bloom's Taxonomy as refined by (Anderson & Krathwolh, 2002: 215) consist of six cognitive domains, namely: remember, understand, apply, analyze, evaluate, and create.
Gunawan in Fanani (2013) explains that higher order thinking skills, namely the process of thinking by processing existing ideas with specific techniques to provide understanding at a high level. Improving students' higher-order thinking skills can be given through HOTS-based question stimulus (Istiyoo, Mardapi, Suparmo, 2014: 3 in Siti Fatimah, 2020: 318). Meanwhile, according to Awaliyah, 2018: 47 explains that the skills of students in working on and answering questions with a reasoning process, being able to solve problems, analyze, reflect and argue are the efforts of the HOTS-based assessment instrument.

The Higher Order Thinking Skills (HOTS) question instrument is a question that tests the level of higher order thinking skills, namely the ability not only to remember, restate, or refer without processing (Dirjendikdasmen, 2017: 3). The characteristics of the HOTS questions according to the Directorate General of Primary and Secondary Education (2017: 4) are divided into three, namely: 1) measuring higher-order thinking skills, 2) based on contextual problems and 3) using various forms of questions.

The first characteristic of HOTS questions, HOTS questions measure the level of higher-order thinking skills, including according to the dimensions of cognitive processes in Bloom's Taxonomy, namely the ability to analyze, evaluate and create. According to Krathwohl (in Lewy, Zulkardi, & Aisyah, 2009: 16), indicators for measuring higher order thinking skills include 3 descriptors of each ability, namely the ability to analyze, evaluate and create.

The second characteristic of the HOTS questions, the preparation of the HOTS questions, explains that questions that are included in Higher Order Thinking have the following characteristics: 1) transfer from one concept to another, 2) process and apply information, 3) seek links from different kinds of information, 4) use information to solve problems, and 5) examine ideas and information critically. (Wiwik Setiawati et al, 2019:139)

In Gregory Schraw and Danile R. Robinson, 2011:50, Those premises the nature og higher order thingking meant were as follow :
1)higher-order thingking is dificult to define but easy to recognize when it occurs, 2) higher order thingking has always been is to find ways to teach higher order thingking within institutions committed to educating hte entire population, 3) higher order thingking is the hallmark of successful learning at all levels , not only the more advanced.

Method

This research was development research that refers to the type of research and development (R & D). This development model considered suitable because it develops an instrument for assessing the higher order thinking skills of junior high school students in science subjects. In essence, R & D research was an effort to develop an effective product to use, and not to test theory. Specifically, R&D research in education was the process used to develop and validate educational products. The steps of the R & D research process consist of studying research findings related to the product to be developed, developing a product based on those findings, conducting testing, and revising to correct deficiencies found in the field. In addition, R & D research also has advantages in terms of detailed and practical work procedures.

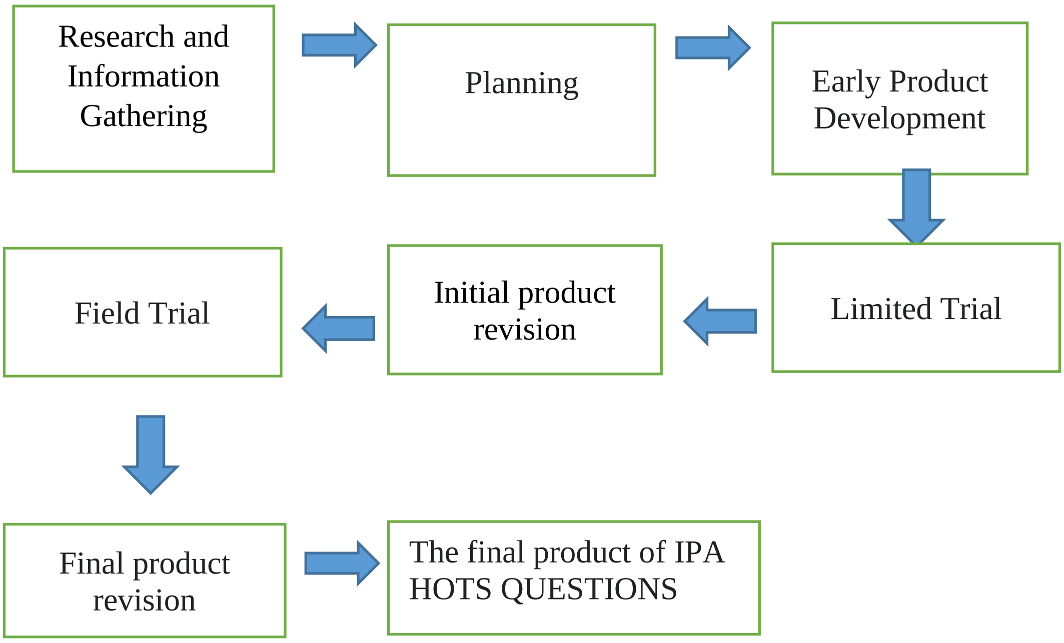

Figure 1. Stages of research and development

The initial stage of research and information gathering was collected information and data as the basis for developing the HOTS assessment instrument. This was done through pre-research activities to find out the facts in the field and also through literature review to support development research. The next activity is planning. Activities carried out in planning are formulating research objectives to be achieved and the capabilities needed to develop products. Conduct material analysis in accordance with the SMP/MTs curriculum. The next stage was the initial product development which is the process of making the initial product design in the form of a HOTS assessment instrument. The results of this initial design resulted in a draft of 1 product. After the draft 1 product is ready, it is validated by an expert. Validation was carried out by science experts to obtain input for improving the HOTS assessment instrument and to find out whether the developed instrument was feasible and met the valid criteria before being tested. After the validation process, product I revision was carried out, based on expert input to produce draft 2 of the HOTS assessment instrument product;

The next stage was to conduct a limited trial of the product. The goal is to determine the legibility of the product being developed. The legibility aspect in question is whether students understand the intent of each item in the instrument and to ensure sufficient time in working on the questions. Furthermore, a product revision was carried out from the results of a limited trial (revision of product II). This stage produced a draft of 40 HOTS assessment instrument products. Furthermore, field trials were carried out to determine the quality of the development product. The quality in question is from the results of estimating its reliability, analyzing the distinguishing power and level of difficulty that can be shown from the HOTS assessment instrument. This trial consisted of 140 students. The final stage was the revision of the final product and the assessment of the final product of the development results to decide whether or not the assessment instrument is used to measure students' HOTS abilities.

Data Collection Techniques and Instruments

The types of data obtained from this research and development were quantitative and qualitative data. This data provided an overview of the quality of the products developed, in this case the HOTS IPA class IX instrument. Qualitative data is obtained from validator input as well as from converting quantitative data to the specified category. While quantitative data obtained from the results of field trials. The research instrument used consisted of instruments to determine valid criteria and to determine reliability. The instrument used to measure the validity is the item conformity validation sheet with the indicator. Validation is also reviewed from three aspects, namely material, construction and language. The instrument for measuring reliability was a set of multiple-choice questions. The HOTS instrument was tested individually and the results of the analysis were quantitative to determine the reliability, discriminating power, level of difficulty and effectiveness of distractors.

Data analysis technique.

According to Retnawati. 2016: 18, content validity was determined using expert agreement. The agreement of experts in the field of study or often referred to as the measured domain determines the level of content validity (content-related). This was because the measurement instrument, for example in the form of a test or questionnaire, is proven valid if the expert believes that the instrument measures mastery of the abilities defined in the domain or also the psychological construct being measured. To find out this agreement, validity indices can be used, including the index proposed by Aiken (1980; 1985; Kumaidi, 2014). The item validity index proposed by Aiken is formulated as follows:

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

where V is the index of rater agreement regarding item validity; s score assigned by each rater minus the lowest score in the category used (s = r – lo, where r = score in the rater's choice category and lo the lowest score in the scoring category); n number of raters; and c the number of categories that the rater can choose from.

A test is said to be reliable or steady if it is tested several times and gives relatively the same results.

$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{S^2-\Sigma pq}{S^2}\right)$	information : r11: Overall test reliability p : Proportion of subjects who answered the item correctly q : Proportion of subjects who answered the item incorrectly (q = 1 – p) n : Number of items s : Standard deviation of the test,

(Arikunto, 2013: 101)

After the analysis results are seen from the reliability of the questions, the results of the reliability calculations are consulted into the interpretation of the reliability values as follows:

Table 1. Reliability Criteria

The value of r	Interpretation
0,80 – 1,00	Very high
0,60 – 0,79	Tall
0,40 – 0,59	Enough
0,20 – 0,39	Low
0,00 – 0,19	Very low

The formula for the level of difficulty (P) is:

$P = \frac{B}{JS}$	Information P : Index of difficulty/level of difficulty B : The number of students who answered the question correctly JS : The total number of students in the test.

(Arikunto,2013:208)

According to Wayan Nurkancana (2002:161) suggests about the classification of the level of difficulty of the questions, namely:

• if the value of p (percentage level of difficulty of the problem) = 0.81 – 1.00 the item is straightforward,
• if the value of p = 0.61-0.80 easy questions,
• if the value of p = 0.41 – 0.60 moderate items,
• if the value of p = 0.21 – 0.40 difficult items,
• if the value of p = 0.00 – 0.20 items very difficult

The objective form test in calculating discriminatory power can be done using the following formula:

$D = P_A - P_B$ $P_B = \frac{B_B}{J_B}$	Information : D : The sought-after distinguishing power BA : Upper limit JA : Amount of lower limit JB : Number of lower limit : Proportion of true top (remember P, as difficulty index) : Proportion of the lower group who answered correctly

(Arikunto2013:214)

After getting the results of discriminating power, the results are clarified based on the quality of the questions. This is done to make it easier to determine the quality of the questions that have been made according to the results of these calculations. Then Arikunto (2013: 218) clarified the questions in accordance with the results of the calculations above, namely as follows:

Table 2. Clarification of Distinguishing Power

Distinguishing Power Index	Category
0,00 – 0,19	Bad
0,20 – 0,39	Enough
0,40 – 0,69	Well
0,70 – 1,00	Very well
Negatif	Everything is not good, so all questions that have a D value should be discarded

The distractor index is calculated using the formula:

$IP = \frac{P}{(N-B)/(n-1)} \times 100\%$	Information : IP: Detractor index Q: Number of students who chose distractors N: Number of students who took the test B: Number of students who answered correctly N: Number of alternative answers (option)1: Fixed number.

(Arifin, 2016: 270)

Table 3 Criteria for Assessment of the Effectiveness of Detractors

Working Answer	Criteria
4 answer options	Very good
3 answer options	Well
2 answer options	Pretty good
1 answer option	Not good
None of the answer options work	Not good

Results and Discussion

The result of the development in this study is the HOTS instrument in the form of multiple-choice questions consisting of 24 science questions for class IX. The product development was in the form of a HOTS instrument that passes through two stages of assessment, namely from the results of validation and testing. The expert validation involved 5 science experts consisting of one former supervisor of SMP in Magelang Regency, two Principles of a State Junior High School in Magelang Regency and two Lecturers from the Bachelor Wiyata University

Tamansiswa Yogyakarta. The limited trial involved 10 grade IX students of SMP Negeri 1 Sawangan and the field trial involved 140 grade IX students of SMP Negeri 1 Sawangan, Magelang Regency. The core process carried out in this development research was compilation of a draft of the HOTS Science instrument, expert validation, revision 1, limited trial, revision II and field trials, until the final product of the HOTS Science Class IX instrument is ready and suitable for use. The initial draft HOTS instrument consisted of 40 multiple choice questions on the material of the human reproductive system and the reproductive system of plants and animals.

Results Validation

The item validity index with the Aiken V index is an index of rater agreement on the suitability of the item (or whether or not the item is appropriate) with the indicator you want to measure using the item. If applied to a measuring instrument, according to a rater, n can be replaced by m (the number of items in one instrument). The value of this V index ranges from 0-1. From the results of the calculation of index V, an item or device can be categorized based on its index. If the index is less than or equal to 0.4 it is said to be less valid, 0.4-0.8 is said to be moderately valid, and if it is more significant than 0.8, it is said to be very valid.

Based on the results of the analysis using the Aiken's V formula, it shows that the items with moderate validity are item number 1 while items number 2 to 40 are said to be very valid. With these results, all multiple-choice questions, totaling 40 questions, are feasible to be used for testing. Based on the results of a qualitative study on 40 items, it is overall good. The average material aspect is at 89%, which means it shows a good category because the grid and the question indicators are in accordance with what is being measured. In the aspect of construction and language the average is at 97% which means it also shows a good category. Thus, 40 questions are feasible and suitable to be used as test instruments.

Product Trial Results

Limited trials were conducted to determine the readability of the HOTS instrument product. Readability in this case is that students can understand the questions contained in the instrument and measure the time that might be used in field trials. This limited trial involved 10 grade IX students of SMP Negeri 1 Sawangan Magelang. From the results of the readability test, it is known that it takes 120 minutes to work on the HOTS Science questions. Some editorial questions have been revised to make it easier for students to understand and most of the language used is easy to understand, the sentences used are unambiguous and look attractive so that the questions can be used.

After revisions were made based on the results of a limited trial, the HOTS instrument was field tested involving 140 grade IX students of SMP Negeri 1 Sawangan Magelang. The results of the field trials were analyzed using Anatses 4.09 to see the validity, reliability, discriminating power, level of difficulty and effectiveness of distractors.
Validity in this study, calculated using Anates 4.09. From the results of the analysis of 40 multiple-choice questions HOTS science maple class IX used for testing, it is shown that there are 24 valid (significant/very significant) questions (60%), there are 16 items that are not valid (40%)). If the 40 multiple-choice HOTS science subjects for class IX are distributed based on their validity index, the results will be obtained as shown in table 14 below.

Table 4. Distribution of Question Item Validity

No	Validity Index	No. Question Points	Amount	(%)
1	If the index value is above or equal to 0.35, then the question is declared valid	2,4,5,7,10,11,13,14.17,21,23,24,25,26,28,29,30,31,32,33,36,37,39,40	24	60 %
2	If the index value is less than 0.35 then the question is declared invalid	1,3,6,8,9,12,15,16,18,19,20,22,27,34,35,38	16	40%

The results of the analysis can be seen through the following pie chart:

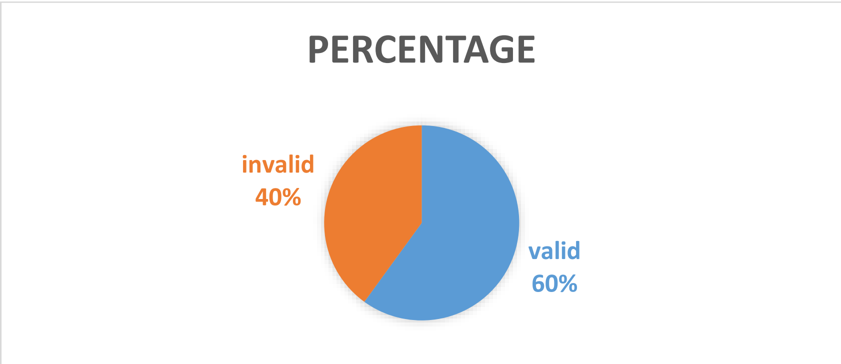

Figure 2. Pie Diagram Percentage of Item Validity

From the results of the analysis of the Anates 4.0.9 program, the data obtained from 40 multiple choice items that have been analyzed, it can be seen that the reliability of the test is 0.61. Based on these results, it can be said that the reliability of the test for HOTS questions for science subjects class IX in this study, has a high interpretation. This is because the test reliability of 0.61 in this research trial is included in the test reliability of 0.60 - 0.79 which refers to the reliability criteria formula. Moreover, after analyzing 24 valid questions, the reliability of the test became 0.85 in the very high category.

Based on the results of the analysis through the Anates 4.0.9 Program on 40 multiple-choice HOTS questions for Science Class IX subjects that were tested, it can be obtained information on the level of difficulty that there are as many as 1 question in the specific category, 11 questions in the easy category, 20 questions in the easy category. medium, five questions in the difficult category, and three in the tricky category. If distributed based on the index of difficulty level, the results can be seen as follows.

Table 5. Distribution of Item Difficulty Levels

No.	Category	No. Question	amount	(%)
1	Very difficult (0,00–0,02)	1,18,35	3	7,5 %
2	Hard (0,21 - 0,40)	22, 12, 19, 27, 34	5	12,5 %
3	Currently (0,41 -0,60)	4, 6, 7, 9,11,13,14,15,16,17, 20, 22, 23, 24, 26, 29, 33, 38, 39, 40	20	50 %
4	Easy (0,61- 0,80)	2, 5, 10, 21, 25, 28, 30,31, 32, 36, 37	11	27,5 %
5	Very easy (0,81 - 1)	3	1	2,5 %

If the results of the analysis of the multiple-choice HOT items for the Science Subjects of class IX are converted into pie charts, the results can be seen in the following diagram.

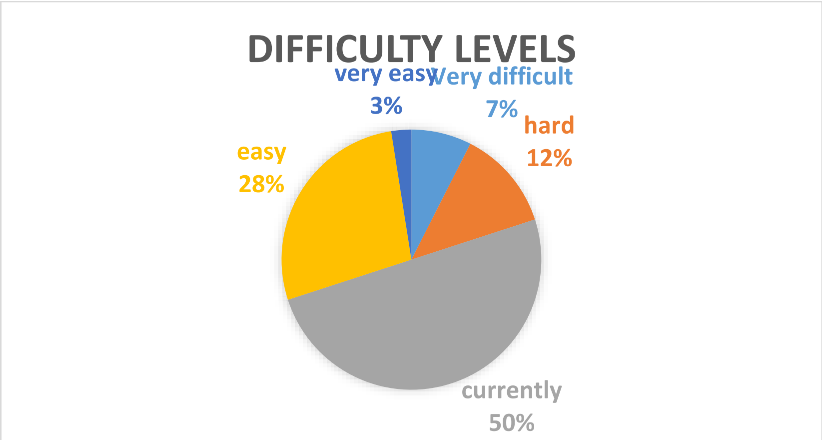

Figure 3. Pie Diagram of the Percentage of Item Difficulty Levels

The discriminatory power obtained based on the results of the Anates Program 4.0.9 analysis and can be obtained information that from a total of 40 multiple choice questions tested; there are four items (10%) in the wrong category, ten items (25%) in the excellent category, 17 items (42.5%) in the excellent category, 1 item (2.5%) in the outstanding category, and eight items (20%) in the harmful category. Based on the results of this study, it was found that there were five criteria that emerged, namely the items in the categories of evil, sufficient, good, excellent, and harmful. However, this does not affect the overall results because the numbers that appear are the result of the program directly. If
distributed based on the index of discriminating power, the results can be seen in the following table.

Table 6. Distinguishing Power Distribution of Items

No.	Distinguishing Power Index	Category	No. Question	amount	(%)
1	0,00 – 0,19	Bad	3, 6, 9, 35	4	10 %
2	0,20 – 0,39	Enough	7, 15, 17, 20, 22, 23, 24, 29, 38, 40	10	25 %
3	0,40 – 0,69	Well	2,5, 10, 11, 13, 14, 21, 25, 26, 28, 30, 31, 32, 33, 36, 37,39	17	42,5 %
4	0,70 – 1,00	Very well	4	1	2,5 %
5	Negatif	Everything is not good.	1,8,12,16, 18, 19, 27, 34	8	20 %

If the results of the item analysis in terms of discriminating power are converted into pie charts, the results can be seen below.

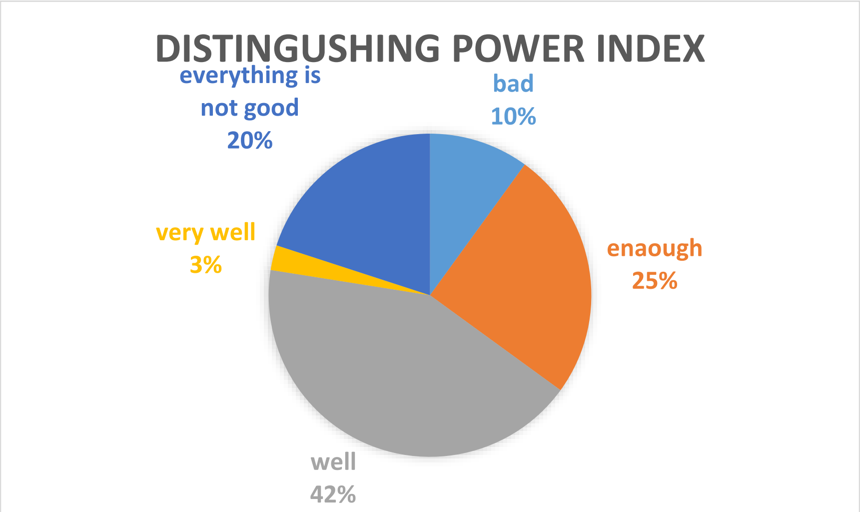

Figure 4. Pie Diagram of the Percentage of Distinguishing Power of Items

In terms of the effectiveness of the distractors in the trial of this study, it showed that from a total of 40 multiple-choice questions, there were 16 items (40%) in the perfect category (++), three items (7.5%) in the excellent category (+), ten items (17%) in the poor category (-), two items (7%) in the wrong category (- -), and three items (10%) in the very bad category (- - -). If distributed based on the Detractor Effectiveness Index, the results can be seen in the following table.

Table 7. Distribution of the Effectiveness of Distracting Items

No.	Information	Category	No. Question	amount	(%)
1	++	Very good	1,4, 7, 10,11,17, 18, 22, 24, 29, 31, 32, 34, 36, 37, 39	16	40 %
2	+	Well	21, 30, 33	3	7,5 %
3	-	Not good	5, 12, 13, 14, 16, 23, 27, 28, 35, 38	11	27,5 %
4	- -	bad	5	1	2,5 %

In terms of the effectiveness of the distractors, they are converted into pie charts; the results can be seen in the following diagram.

In terms of the effectiveness of the distractors, they are converted into pie charts, and the results can be seen in the following diagram.

In terms of the effectiveness of the distractors, they are converted into pie charts. The results can be seen in the following diagram.

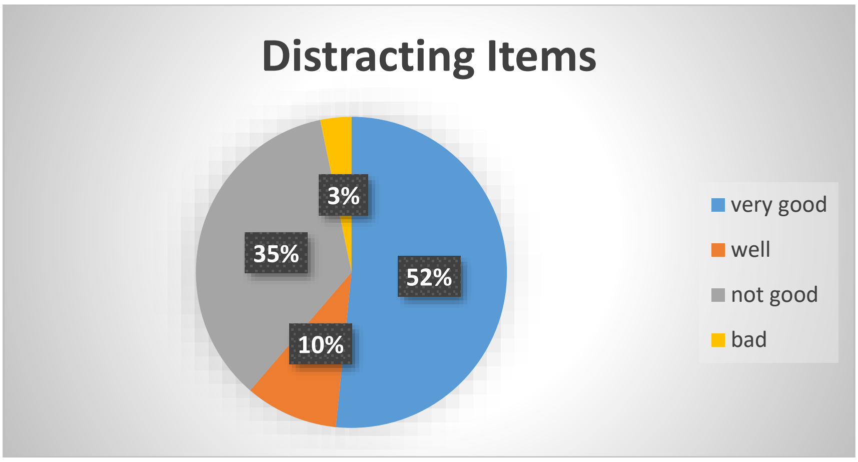

Figure 5. Pie Diagram of the Effectiveness of the Percentage of Item Distractors

Conclusion

Analysis of multiple-choice HOTS questions for IPA Class IX, using the Anates Program Version 4.0.9 in terms of validity, there are 24 valid questions, reliability of 0.85 in the very high category, the level of difficulty is dominated by 50% of moderate questions, the distinguishing power is as much as 70% of the questions included having adequate discriminating power, and the effectiveness of the distractor as much as 47.5% of the questions were in the very good and good categories. Thus, in this study, the quality HOTS Science Class IX questions consisted of 24 multiple choice questions and could be used to measure the HOTS Science skills of class IX students.

References

Anderson, L.W., dan Krathwohl, D.R. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Addison Wesley Longman, In.

Arifin, Z., & Retnawati, H. (2015). Analisis Instrumen Pengukur Higher Order Thinking Skills (HOTS) Matematika Siswa SMA. Seminar Nasional Matematika Dan Pendidikan Matematika Uny.

Arikunto, Suharsimi. 2013. *Dasar-dasar Evaluasi Pendidikan Edisi 2*. Jakarta:BumiAksara.

Awaliyah, S.2017. Penyusunan Soal HOTS bagi Guru PPKN dan IPS Sekolah Menengah Pertama . Jurnal Praktis dan Dedikasi Sosial

Borg, W. R., & Gall, M. D. (1983). Educational research and introduction. New York, NY: Longman.

Budiman, A., & Jailani, J. (2014). Pengembangan Instrumen Asesmen Higher Order Thinking Skill (Hots) Pada Mata Pelajaran Matematika Smp Kelas Viii Semester 1. Jurnal Riset Pendidikan Matematika

Conklin, W. (2012). Higher-Order Thinking Skills to Develop 21st Century Learners. Shell Education. https://doi.org/10.1017/CBO9781107415324.004

Direktorat Jenderal Guru dan Tenaga Kependidikan. 2019. Buku Penilaian Berorientasi Higher Order Thinking Skills. Diterbitkan oleh Direktorat Jenderal Guru dan Tenaga Kependidikan Kementerian Pendidikan dan Kebudayaan : Jakarta

I Wayan Widana, 2017. Modul Penyusunan Soal Higher Order Thinking Skills (HOTS). Jakarta : Direktorat Pembinaan SMA Direktorat Jenderal Pendidikan Dasar dan Menengah Departemen Pendidikan dan Kebudayaan

Ernawati, L. 2017. Pengembangan High Order Thinking (Hot) Melalui Metode Pembelajaran Mind Banking dalam Pendidikan Agama Islam. 1st International Conference on Islamic Civilization ans Society (ICICS). Diselenggarakan oleh Darul Ulum Islamic Unversity 28.Jakarta: Pusat Penilaian Pendidikan Balitbang-Depdiknas

Fanani,M.Z. 2018. Strategi Pengembangan Soal Higher Order Thingking Skils (HOTS) dalam Kurikulum 2013.E Dudeena Jounal Of Islamic Religous Education.

Gregory Schraw and Danile R. Robinson, 2011. Assessment of Higher Order Thinking Skills. Information Age Publishing.

Heri Retnawati. 2016. Analisis Kuantitatif Instrumen Penelitian.Yogjakarta: Parama Publishing

I Wayan Widana, 2017. Modul Penyusunan Soal Higher Order Thinking Skills (HOTS). Jakarta : Direktorat Pembinaan SMA Direktorat Jenderal Pendidikan Dasar dan Menengah Departemen Pendidikan dan Kebudayaan

Kemendikbud, (2017). Modul Penyusunan Higher Order Thinking Skill (HOTS). Jakarta: Direktort Jenderal Pendidikan Dasar dan Menengah Departeman Pendidikan dan Kebudayaan.

Kemendikbud. 2018. Peraturan Menteri Pendidikan dan Kebudayaan Nomor 69 Tahun 2013 Tentang Kerangka Dasar dan Struktur Kurikulum Sekolah Menengah Atas/Madrasah Aliyah. Jakarta: Kemendibud. Kemendikbud. 2018. Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 37 Tahun 2018 Tentang Perubahan Atas Peraturan Menteri Pendidikan dan Kebudayaan Nomor 24 Tahun 2016 Tentang Kompetensi Inti dan Kompetensi Dasar Pelajaran pada Kurikulum 2013 pada Pendidikan Dasar dan Pendidikan Menengah. Jakarta: Kemendibud.

Permendikbud No 21 tahun 2016 tentang Standar Kompetensi Lulusan

Permendikbud No 37 tahun 2018 tentang KI dan KD Kurikulum 2013 pada Pendidikan Dasar dan Menengah

Siti Zubaidah,dkk.2018. Ilmu Pengetahuan Alam SMP/Mts Kelas IX Semester 1. Surakarta : CV Putra Nugraha
Sudijono, Anas. 2012. *Pengantar Evaluasi Pendidikan.* Jakarta: PTRaja Grafindo Persada.
Sunarti, & Rahmawati, S. 2014. Penilaian dalam Kurikulum 2013. Yogjyakarta: Andi Offset
Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional.
Wiwik Setiawati dkk. 2019. Buku Penilaian Berorientasi Higher Order Thingking Skills. Direktorat Jendral Guru dan Tenaga Kependidikan Kementerian Pendidikan dan Kebudayaan