Vol 12 (1), 2023, 13-29 DOI: 10.23960/jppk.v12.i1.2023.02
Jurnal Pendidikan dan Pembelajaran Kimia
e-ISSN: 2714-9595| p-ISSN 2302-1772
http://jurnal.fkip.unila.ac.id/index.php/JPK/index

Development of an Evaluation Instrument to Measure Higher-Order Thinking Skills in Acid-Base Topic

Salsabila Hirza[1]*, Zainuddin Muchtar[2], Ani Sutiani[3], Ratu Evina Dibyantini[4], Marudut Sinaga[5]

1,2,3,4,5 Chemistry Education, Faculty of Mathematics and Natural Science, State University of Medan, Jl. Willem Iskandar, Pasar V, Medan, Indonesia.

*Corresponding e-mail: salsabila27b@gmail.com

Received: *April 6th*, 2023 *Accepted : April: 15th*, 2023 *Online Published:* April 25th, 2023

Abstract: Development of an Evaluation Instrument to Measure Higher Order Thinking Skills in Acid-Base Topic. The study aims to develop an evaluation instrument for higher order thinking skills in acid-base material and determine the feasibility of the developed evaluation instrument. This study uses a Research and Development (R&D) design with a 4D development model. The sample in the study was 36 students of class XI A at SMA Negeri 2 Percut Sei Tuan. The results of the research give the final product in the form of 25 good questions from the initial product design of 40 items. The item is said to be good because it is valid and has a reliability value of 0.815 with a very high category. Discriminating power with a percentage of 94% in the good category, and 47% difficulty level of the questions in the medium category. Students' high-level thinking skills in acid-base material reach 42% in the good category. This is shown from the percentage of answers on cognitive level questions analyzing (C4) of 36%, evaluating (C5) of 52%, and creating (C6) of 12%. Through some of these feasibility tests, it can be concluded that the evaluation instrument developed is good for measuring students' higher-order thinking skills in acid-base material.

Keywords: Evaluation Instrument, Higher Order Thinking Ability, Acids and Bases.

Abstrak: Pengembangan Instrumen Evaluasi untuk Mengukur Keterampilan Berpikir Tingkat Tinggi pada Materi Asam-Basa. *Penelitian ini bertujuan untuk mengembangkan instrumen evaluasi keterampilan berpikir tingkat tinggi pada materi asam basa dan mengetahui kelayakan instrumen evaluasi yang dikembangkan. Penelitian ini menggunakan desain Research and Development (R&D) dengan model pengembangan 4D. Sampel dalam penelitian yaitu 36 siswa kelas XI A di SMA Negeri 2 Percut Sei Tuan. Hasil penelitian memberikan produk akhir berupa 25 soal yang baik dari rancangan awal produk 40 butir. Butir dikatakan baik karena telah valid dan memiliki nilai reliabilitas 0,815 dengan kategori sangat tinggi. Daya pembeda dengan persentase 94% kategori baik, dan 47% tingkat kesukaran soal pada kategori sedang. Keterampilan berpikir tingkat tinggi siswa pada materi asam basa mencapai 42% dalam kategori baik. Hal ini ditunjukkan dari persentase jawaban pada soal level kognitif menganalisis (C4) sebesar 36%, mengevaluasi (C5) sebesar 52%, dan mengkreasi (C6) sebesar 12%. Melalui beberapa uji kelayakan tersebut, dapat disimpulkan bahwa instrumen evaluasi yang dikembangkan baik digunakan untuk mengukur keterampilan berpikir tingkat tinggi peserta didik pada materi asam basa.*

Kata kunci: *Instrumen Evaluasi, Keterampilan Berpikir Tingkat Tinggi, Asam Basa.*

▪ INTRODUCTION

In the 21st century, human resources need three main skills: critical reasoning, creative thinking, and problem solving. These three skills are known as higher-order thinking skills (HOTS). Sani (2019) said that the rapid development of information and technology brings difficulties and the problems that humans will face in the twenty-first century are increasingly complicated. Therefore, it is important to prepare provisions for the younger generation to think creatively and critically and make decisions to solve problems (Saraswati dan Agustika, 2020).

In the findings of the PISA study, the 74th position out of 79 countries occupied by Indonesia on secondary education systems worldwide in 2018 was associated with science education in 2019. In other words, in terms of countries that registered, Indonesia is ranked sixth (OECD, 2019). This is a very concerning condition. It is unfortunate that even though there are many human resources in Indonesia, education has not been able to improve the standard of these resources. Many variables hinder the progress of Indonesia's education system, including the country's very poor education standards compared to other countries (Kurniawati, 2022). Chemistry learning is included in the scope of science learning, so that it also supports the achievement of students' science literacy skills (Anggreani et al, 2016).

After studying science, students acquire the generic skills of science, which is the ability to reason and act according to scientific information. In the study of scientific concepts and ways of solving problems used generic skills of science. This skill is one of the main abilities needed by humans in the 21st century. Basicly, the way of thinking and doing in studying the concept of science and problem solving is the same (following the triangular principle of nature assessment), therefore generic competence exists (Saputra, 2016).

Chemistry is part of science, studying chemistry aims to find out about events experienced in the real world. Therefore, learning chemistry involves developing cognitive mastery in the form of theory and logic and thinking skills (Nurkholik & Yonata, 2020). Therefore, to solve problems involving theories, concepts, laws, and facts, higher order thinking skills (HOTS) are required when studying chemistry (Danggus, 2014). However, the majority of students view chemistry as a challenging topic. This is because of how abstract and complicated the idea of chemistry is. Students in chemistry must switch representations between macroscopic, submicroscopic (molecular), and symbolic (iconic).

The implementation of the K-13 curriculum focuses on improving two main components, namely content standards and evaluation standards. This is one of the government's efforts to improve the quality of higher-order thinking skills (HOTS) students (Ministry of Education and Culture, 2017). Content standards are created so that students can think critically when they receive different types of information, think imaginatively when they use their knowledge to solve problems, and make choices in challenging circumstances (Saputra, 2016). The global standard assessment model is modified to apply assessment standards, and the evaluation method emphasizes higher-order thinking skills (HOTS) (Kemendikbud, 2017).

Based on Bloom and Brookhart's (2010) taxonomy, there are two perspectives on the capacity for higher-order thinking. The ability to knowing-C1, understanding-C2, applying-C3, analyzing-C4, evaluating-C5, and creating-C6 is a dimension of the thinking process in Bloom's taxonomy. HOTS achievements are shown in categories C4, C5, and C6 based on Bloom's taxonomy. By utilizing this measure of ability, one can

determine one's capacity for higher-order thinking (Anderson and Krathwohl, 2002). According to Brookhart (2010), the HOTS evaluation instrument includes the ability to think critically, solve problems or find solutions, and think creatively.

According to Kusuma (2017), using HOTS evaluation questions as an alternative to train and measure students' HOTS levels is a good idea for teachers. However, the fact is that the use of HOTS evaluation questions is still rarely used in assessment. This is because when making HOTS questions, teachers must be able to display different information using stimuli, such as information from real life that can be found in text, images, graphs, tables, etc. (Merta, Lestari, & Setiadi, 2019). To make the questions more effective, the stages of creating the HOTS question item should be carefully followed. This includes choosing basic competency that can be converted into HOTS items, making question grids, determining relevant and interesting stimuli, making question items according to the question grid, and making assessments (Fanani, 2018).

The main problem in the field is that teachers lack a thorough understanding of how to prepare and make HOTS questions (Salirawati, 2017). The evaluation questions used for assessment only evaluate thinking skills at the memorization stage, or lower-order thinking skills (LOTS), according to the findings of an interview at SMA Negeri 3 Binjai. Teachers believe that the topic of HOTS evaluation is challenging and requires special skills in making it, so it is still relatively rarely used. Therefore, students' HOTS was not measurable and trained. Given these circumstances, the creation of assessment tools to measure students' higher-order thinking skills needs to be developed.

According to the Ministry of Education and Culture (2017), HOTS questions are a useful assessment tool for assessing thinking skills that go beyond simple memory, rearranging, or referring without considering processing. In the context of evaluation, HOTS questions measure students' capacity for concept transfer, information application, searching for relationships among a variety of varying pieces of information, problem solving, and critical analysis of concepts and information.

The topic of acid-base chemistry was chosen for the development of the HOTS problem. Stieff and Wilensky (2003) explain that the topic of solid acids and bases contains concepts and requires the integration of understanding of many areas of introductory chemistry. The acid-base topic also prioritizes two elements, namely conceptual and algorithmic. In calculating pH or pOH, identifying Ka and Kb, and the percent ionization of acid-base solutions, algorithms are used. While conceptual contains explanations for various acid-base phenomena that occur in life. (Drechsler & Schmidt, 2005).

Based on this information, researchers are interested in conducting studies on the development of evaluation instruments to measure higher-order thinking skills in acid-base materials. The objectives of this study are: 1) developing an evaluation instrument that can measure higher-order thinking skills on acid-base material; 2) knowing the feasibility of an evaluation instrument in measuring higher-order thinking skills on acid-base material; 3) knowing the higher-order thinking skills of grade XI A SMA Negeri 2 Percut Sei Tuan students in completing an evaluation instrument to measure higher-order thinking skills on acid-base material.

▪ METHODE

The research design used in this study is research and development (R&D). This study used quantitative approaches and evaluation instruments to measure HOTS on the acid-base material of the developed product. A 4D development model was used in this

study. Thiagarajan, et al (1974), stated that the 4D model consists of 4 stages, namely define, design, develop, and disseminate.

1) Define Stage

The define stage is a stage that includes the activities of defining the product developed along with its definition and specifications. Early final analysis is in the form of problem analysis to alternative analysis that can be developed to solve the problem, student analysis, task analysis, and concept analysis. This stage is carried out by conducting literature studies in books, journals, and other references.

2) Design Stage

The design stage is the stage of designing the evaluation instrument grid to be developed. This stage is carried out after analyzing the learning material and determining the goals to be achieved. At this stage, question indicators are designed on each pound of purpose and also determine the form and cognitive level of the test instrument to be developed. Then as a result of this stage, an initial draught of the evaluation instrument is obtained.

3) Develop Stage

The develop stage is the stage of developing evaluation instrument products based on the designed question grid. After that, the evaluation instrument product was validated by experts and then a revision of the evaluation instrument was carried out based on input from validators. After the evaluation instrument is revised, feasibility testing is carried out on the evaluation instrument by implementing it in small-scale trials.

4) Dessiminate Stage

This disseminate stage is the stage of using learning tools that have been developed.

This research was conducted at SMA Negeri 2 Percut Sei Tuan which is located at Jalan Pendidikan Pasar XII, Bandar Klippa Village, Percut Sei Tuan District, Deli Serdang Regency, North Sumatra Province. The study was conducted from December 2022 to January 2023. The sample used was 36 students from class XI A SMA Negeri 2 Percut Sei Tuan selected using random sampling technique (Arikunto, 2010).

▪ RESULTS AND DISCUSSION

Result

1) Development Phase of Evaluation Instruments to Measure Higher Order Thinking Skills on Acid-Base

The results of the development process research were obtained based on the success at each stage of the 4D development process. The results obtained at each stage of the development process, namely:

A. Define Stage

This stage includes initial final analysis, learner analysis, task analysis, concept analysis, and goal analysis.

a. Initial Final Analysis

The initial final analysis aims to find alternative solutions. The initial final analysis seeks to identify the difficulties encountered in studying chemistry, especially in the field of acid-base materials. The results of the final initial analysis indicate that an evaluation instrument to measure HOTS on acid-base material needs to be developed. A common problem encountered is that test questions used in schools are still classified as LOTS criteria, namely at the cognitive level C1 (remembering), C2 (understanding), C3 (applying) according to Bloom's taxonomy, so students are only used to doing low

category questions. On the other hand, teachers are not used to making HOTS questions due to limited time and the need to achieve learning objectives. So that from these problems, a HOTS evaluation instrumen that has good validity and reliability is needed.

b. Student Analysis

Student analysis activities are focused on grade XI high school students as test subjects because grade XI students are included in the stage of formal operations in Piaget's theory. There are those who can think abstractly, logically, and more idealistically between the ages of 11 and adulthood or adolescence. With this ability, students can develop their thinking skill to solve problems and draw conclusions in a structured manner. Students who initially believe that they would only be able to understand concepts should also be encouraged to be able to connect some of these concepts into ideas, ideas, and richness to raise their level of thinking. Later, creativity and innovation will also come up with ideas, and resources. This can be done one of them by getting used to solving complex problems such as HOTS questions.

c. Task Analysis

To select the material to be used on the evaluation instrument to measure HOTS, task analysis is required. The result of the task analysis is the preparation of learning indicators in accordance with the 2013 curriculum syllabus of acid-base material. Acid-base material is contained in Basic Competencies 3.10 and 4.10 with sub-materials on the development of acid and base concepts, acid-base indicators and acidity degrees.

d. Concept Analysis

With the help of concept analysis, evaluation instruments for measuring HOTS in acid-base materials can be created by identifying, collecting, and linking existing concepts.

Researchers conducted concept analysis by identifying acid-base concepts from each existing sub-material. After that it was arranged into a collection of sub-sub-matter where in the sub-material the development of the acid-base concept consisted of Arrhenius theory, Bronsted-Lowry theory and Lewis's theory. In the acid-base indicator sub-material consists of universal indicators, litmus paper and Natural indicators. While the degree of acidity consists of a strong acid pH, strong base, weak acid, and weak base. Furthermore, the concepts that have been compiled are associated with existing problems such as in everyday life, conceptual, theoretical, practical, and so on.

e. Purpose Analysis

Studies were conducted to determine the formulation of learning objectives to be achieved in learning. The formulation of goals is carried out with the expectation that students show positive behavioral changes both in terms of knowledge, skills, and attitudes.

In this study the objectives are made based on predetermined learning indicators. Learning objectives include all sub-materials contained in acid-base material. Objectives are formulated using the provisions of ABCD or Audience, Behavior, Condition, and Degree.

B. Design Stage

At this stage, a design has been made to develop an evaluation instrument to measure HOTS in acid-base materials. This stage consists of determining the shape of the instrument, preparing the question grid, making answer keys, and designing the instrument.

a. Determination of Instrument Form

At this stage, the form of the HOTS instrument was determined in the form of multiple choice with five answer choices. The use of multiple-choice questions as an educational assessment tool is a trend that is widely used throughout the world, according to Zaman et al, (2010) in Yustika (2014). The use of multiple-choice tests has many additional benefits, such as the ability to cover a wide range of subjects, simple scoring procedures, and the ability of instructors to evaluate test results using computer programs.

b. Preparation of Question Grids

A grid of questions is designed based on the objectives that have been formulated. The question grid is designed by considering the criteria of HOTS questions, namely analyzing, evaluating, and creating. In this case, the researcher is helped by the existence of operational verbs that distinguish between LOTS type questions and HOTS type questions. Khodijah et al, (2021) stated that the purpose of the question grid is to assist in the creation of HOTS question topics. To select basic competencies that can be turned into HOTS questions, cognitive abilities are considered using a grid..

c. Answer Key Generation

The answer key was made with the aim of making it easier for researchers to find the correct answer to each question item. In each question item, there are 5 choice options with 1 correct answer. Answer keys are provided for all 40 developed questions.

d. Instrument Design

The design of the question is adjusted to the material used, namely acid-base. Question making complies with the indicators and learning objectives contained in the grid. The acid-base concept questions consist of 17 questions, acid-base indicators as many as 9 questions, and acidity levels as many as 14 questions.

In addition to making evaluation instruments, researchers also create validation sheets that validators use to assess the feasibility of the instruments developed. Validation is carried out on all question items with assessment aspects, including aspects of material, construction, language, and additional rules. The total statements on the validation sheet are 21 statements.

C. Development Stage

The development phase consists of validator assessment of the instrument used as a basis for revising and refining the instrument. Question validation was carried out by six expert validators, with the help of two chemistry lecturers at Medan State University and four PPG teachers at the same university. Things validated on evaluation instruments to measure higher-order thinking skills include aspects of matter, construction, language, and supplementary rules. The validation stage is carried out by submitting validation sheets, evaluation instruments, question grids, and validator statements. The validation sheet is filled out by checking the assessment column according to the assessment criteria for the evaluation questions, ranging from very bad to very good criteria. The results of the expert assessment determine whether the questions are valid and should be used in schools.

Question items that received suggestions for improvement were revised, after which valid questions could be used for small-scale trials.

D. Dissemination Stage

The dissemination stage was carried out at SMA Negeri 2 Percut Sei Tuan, precisely in class XI A involving 36 students. The questions are distributed in hard copy

form containing 40 acid-base HOTS questions. At this stage, the questions are implemented to students to obtain data for the feasibility test and are used to measure students' HOTS levels on acid-base material. In addition, evaluation instruments are also provided to the school in the form of hard copies and soft copies. Questions can be used and adopted by the school to measure student learning outcomes on acid-base material.

2) Feasibility of Test Instruments

A. Eligibility of Instruments by Expert Validators

The results of the feasibility analysis of evaluation instruments to measure higher-order thinking skills on acid-base material by lecturer validators are presented in the following table.

Table 3.1 Results of Instrument Feasibility Analysis by Lecturer Validators

No	Assessment Aspect	Average		Average
		Validator 1	Validator 2	
1.	Material	4,25	4,63	4,44
2.	Construction	4,50	4,63	4,56
3.	Language	4,25	4,25	4,25
4.	Additional rules	5,00	5,00	5,00
5.	Average	4,50	4,63	4,56
Average Interpretation				Very High
Average Analysis Validation Criteria				Valid

Data in table 3.1, the average feasibility assessment of evaluation instruments by lecturer validator 1 shows results of 4.50 and lecturer validator 2 is 4.63. While the average of each aspect of the assessment shows that in the material aspect, it is 4.44; construction 4.56; language 4.25; and additional rule 5.00. The overall average eligibility assessment was 4.56. It can be concluded that the percentage interpretation is very high, and the problem is feasible to be used to measure HOTS in acid-base topic.

The results of the feasibility analysis of the evaluation instrument for measuring HOTS on acid-base material by PPG teachers are shown in the following table.

Table 3.2 Results of Instrument Feasibility Analysis by PPG Teachers

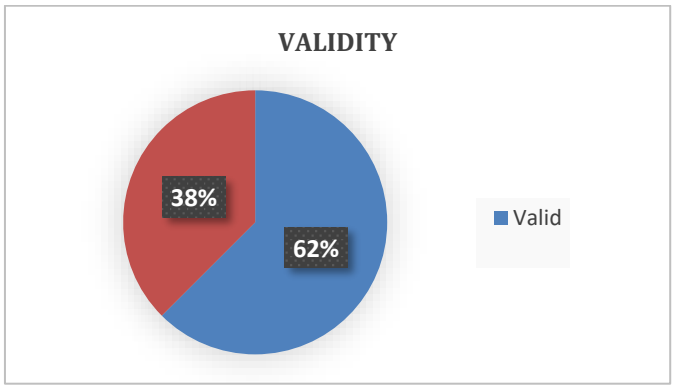
No	Assessment Aspect	Average				Average
		Validator 1	Validator 2	Validator 3	Validator 4	
1.	Material	4,13	4,25	4,25	3,75	4,09
2.	Construction	4,38	4,88	4,38	3,88	4,38
3.	Language	4,50	4,75	5,00	4,00	4,56
4.	Additional rules	4,00	5,00	5,00	4,00	4,50
5.	Average	4,25	4,72	4,66	3,91	4,38
Average Interpretation						Very High
Average Analysis Validation Criteria						Valid

The data in the table showed that the average eligibility of the evaluation instrument as determined by PPG teacher validators showed consecutive scores of 4.25; 4,72; 4,66; and 3.91. The average of each aspect of the assessment was obtained with a value of 4.09 for the material aspect, 4.38 for the construction aspect, 4.56 for the language aspect, 4.50 for the additional rules' aspect. The feasibility analysis of the evaluation instrument yielded a value of 4.38. This means that the evaluation instrument

developed has a very high average interpretation, in other words the evaluation instrument for HOTS on acid-base material is feasible to use.

B. Validity

The validation test results are in the form of quantitative data with an overall average validation value of 0.305. The test instrument is scored by giving a score of 1 for correct answers and a score of 0 for incorrect answers. The data is analyzed with the product moment correlation equation with the help of excel application. The level of significance used is 5% with the question declared valid if the calculation > r_{table}. The percentage of validity of the trial results is briefly presented in the following figure.



Picture 3. 1 Percentage of Validity of Evaluation Instruments

Figure 3.1 shows the percentage of validity test results by obtaining 62% of 40 valid question items (25 question items) and 38% of 40 invalid question items (15 question items). The results of the validity test are presented in the following table.

Table 3.3 Evaluation Instrument Validity Test Results

Criterion	Question Number	Sum	Percentage
Valid	1, 2, 3, 6, 8, 10, 11 ,12, 15, 16, 17, 19, 22, 23, 24, 27, 28, 29, 31, 32, 34, 36, 37, 38, 39	25	62%
Invalid	4, 5, 7, 13, 14, 18, 20, 21, 25, 26, 30, 33, 35, 40	15	18%
Sum		40	100%

C. Test Reliability

The reliability of the evaluation instrument was analyzed using the formula Kuder Richardson 21 (KR-21) which was processed with the help of excel. The reliability calculation result was obtained at 0.815 with a signification level of 5%. This shows that the 25 multiple-choice question items that have been tested have a high level of reliability because the r value ≥ 0.70. Data on the calculation of the reliability of the evaluation instrument can be seen in the following table.

Table 3. 4 Evaluation Instrument Reliability Test Results

No	Number of Items	Reliability	Information
1.	25	0,815	Very high reliability

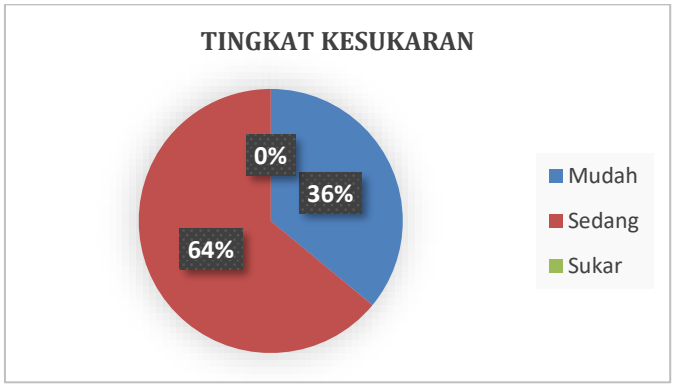
D. Test Difficulty Level

To identify categories of question items that fall into the easy, medium, and difficult ranges, a difficulty level analysis is performed. A good question is a medium-category question. The greater the P value, the easier the question item, on the contrary, the lower the P value, the more difficult the question item. The question qualifies if the price of P ranges from 0.20-0.80 where if P < 0.20 means the question is too difficult and if P > 0.80 means the question is too easy. The difficulty calculation data is presented in the following table.

Table 3.5 Results of the Difficulty Level Analysis of the Evaluation Instrument

Category	Question Number	Sum
Easy	1, 2, 4, 5, 12, 17, 19, 20, 23	9
Keep	2, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 18, 21, 22, 24, 25	16
Difficult	0	0
Sum		25

The data in table 3.5, shows the results that there are 9 questions with easy categories, this shows that almost all students are able to respond to the questions. Furthermore, there are 16 questions that fall into the medium category, meaning that some students can answer the questions correctly. In the evaluation instrument there are no difficult question items, this shows that the evaluation instrument belongs to the group of questions with a good level of difficulty.



Picture 3.2 Percentage of Difficulty Level of Evaluation Instrument

E. Test Differentiating Power

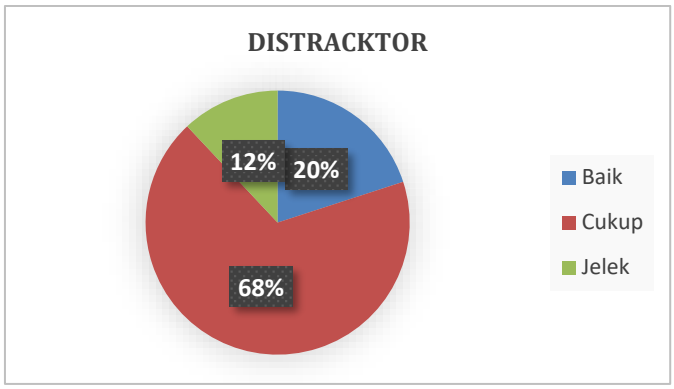
If each test item has a differentiating power of at least 0.2, then it is said to have a good differentiating power. The results of the evaluation instrument trial to students are used to determine the value of the differentiating power. Excel applications are used to process the data obtained. The results of the difference power analysis can be seen in the following table.

Table 3.6 Results of Power Analysis of Difference Evaluation Instruments

Criterion	Question Number	Sum	Percentage
Good	3, 4, 13, 22, 23	5	20%
Enough	2, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 18, 19, 20, 21, 24, 25	17	68%

Bad	1, 11, 17	3	12%

The data in table 3.6 showed the results that 5 questions were included in the "good" group, 17 questions were included in "sufficient," and 3 questions were included in "bad". It can be concluded that there are 22 questions that have good discriminating power, meaning that the questions can distinguish high-ability students and low-ability students, while 3 questions have poor differentiation. The following diagram shows the percentage of the difference power of the problem.



Picture 3.3 Percentage of Differentiating Power of Evaluation Instruments

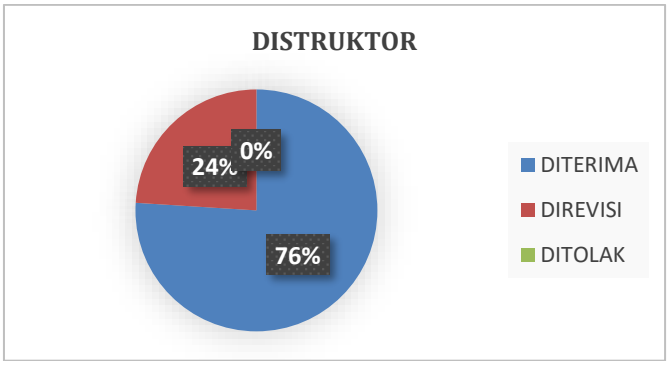
F. Distractor Effectiveness

Analysis of the effectiveness of the distractor is carried out by processing the data of student answers manually with the help of an excel application. To determine whether the deceiver of each question can operate effectively, a distraction effectiveness test is conducted. The distractor is said to work properly if it has been selected by as many as 5% of the test takers. The number of distractors developed in the evaluation instrument to measure HOTS is 125 options from 25 questions. The results of the analysis of the effectiveness of the distractor are presented in the following table.

Table 3.6 Distractor Effectiveness Analysis Results

Criterion	Question Number	Sum	Percentage
Accepted	1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 14, 15, 16, 18, 21, 22, 23, 24, 25	19	76%
Revision	9, 10, 13, 17, 19, 20	6	24%

The data in table 3.6 shows that 19 questions included in the deceptive effectiveness group are accepted meaning that the deceiver can operate well, while there are 6 questions in the revision distractor effectiveness group, meaning that the deceiver has not operated properly and revision is needed before being used.



Picture 3.4 Percentage Analysis of the Effectiveness of Evaluation Instrument Deceivers

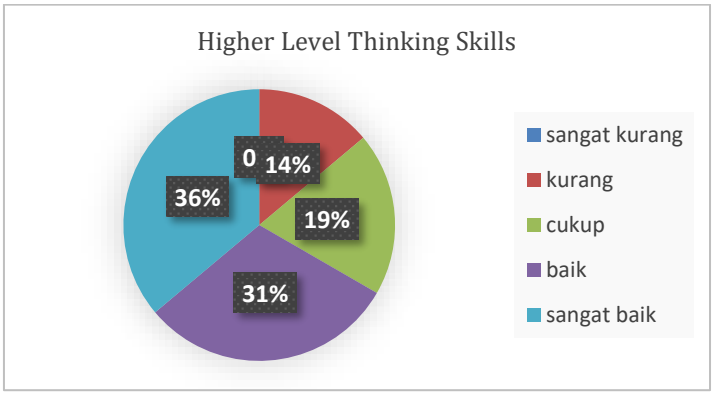
3) Higher Order Thinking Skills

Student HOTS was obtained in a small-scale trial involving 36 grade XI A students at SMA Negeri 2 Percut Sei Tuan. Students do as many as 40 points of evaluation instrument questions on acid-base material with HOTS cognitive levels according to Bloom, namely C4 (analyzing), C5 (evaluating), and C6 (creating). Data was obtained based on the results of student answers processed using excel. The following table shows findings from students' HOTS analysis.

Table 3.7 Analysis Results of Higher Order Thinking Skills

Higher Order Thinking Skills				
No	**Student Grades**	**Criteria**	**Sum**	**Percentage**
1	10 <nilai ≤20	Very Lacking	0	0%
2	20 <nilai ≤40	Less	5,00	14%
3	40 <nilai ≤60	Enough	7,00	19%
4	60 <nilai ≤80	Good	11,00	31%
5	80 <nilai ≤100	Excellent	13,00	36%

The data in table 3.7 shows the results that the highest frequency was among the students with criteria for higher-order thinking skills in the range of 81-100 or very good, which is as much as 36% of the total students. Furthermore, in the range of 61-80 there are 31% of students who have high category abilities. In the sufficient category with a range of 41-60 there are 19% of students and the low category with a range of 21-40 there are 14%. While in the very low category of 0 or there was no students with very low HOTS. The HOTS percentage is depicted in the following diagram.



Picture 3.5 HOTS Rate Analysis Percentage

Students' higher-order thinking skills when reviewed by levels C4 (analysis), C5 (evaluation), C6 (creation) are presented in the following table.

Table 3.8 Percentage of HOTS by cognitive level

Indicators	Sum	Percentage
Analyze (C4)	9	36%
Evaluate (C5)	13	52%
Creating (C6)	3	12%

Based on the data in table 3.8, it was obtained that the largest percentage of student understanding level was in indicator C5 (evaluating) with a value of 52% or totaling 13 questions. In indicator C4 (analyzing) showing results of 36% with a total of 9 questions, and in indicator C6 (creating) results were 12% or with 3 questions.

- **DISCUSSION**

Researchers have completed an evaluation instrument to measure higher-order thinking skills in acid-base matter. This development study is carried out by completing the 4D development stage, namely through the define stage, which is then continued at the design, develop, and disseminate stages. The products produced in this study are 25 evaluation instrument questions made based on Bloom's cognitive level, namely C4 (analyzing), C5 (evaluating), and C6 (creating). With each level, namely 9 C4 questions, 13 C5 questions, and 3 C6 questions. Evaluation instruments have gone through the stages of expert validation and empirical validation, namely validity, reliability, difficulty, differentiation, and distraction tests. According to Alfajri et al, (2019), the device must go through several stages before being considered a finished product, including logical validation by experts and empirical validation, which includes the validity of test items, reliability tests, level of difficulty, distinguishing power, and effectiveness of distractions.

Expert validation was obtained by involving 6 validators, namely 2 validators of chemistry lecturers at Medan State University, and 4 validators of PPG teachers at Medan State University. An evaluation instrument of 40 questions with 5 multiple-choice options. The assessment was carried out using a validation sheet according to BSNP which contained 21 statements with 4 criteria covering aspects of material, construction, language, and additional rules. In addition, validators are also asked to provide advice on each item of the instrument developed. The suggestions given by validators are improvements in terms of sentence use, cognitive level adjustment, and writing systematics. The average assessment by lecturer validators is 4.56 which means it is very high and the questions are worth using. While the average validation by PPG teachers is 4.38 which means very high. Based on expert validation, it can be stated that evaluation instruments to measure higher-order thinking skills on acid-base materials can be used. Questions that have been validated by experts are revised before being used for limited trials in class XI A SMA Negeri 2 Percut Sei Tuan.

The trial was conducted by giving a revised evaluation instrument to 36 students along with an answer sheet to answer the question. 40 HOTS questions made according to acid-base learning objectives are given to students. The problem has included several acid-base sub-materials including the concept of acid-base, acid-base indicators, and acidity degrees. The results of student answers are processed to obtain question validation

data, reliability, level of difficulty, discriminating power, effectiveness, and used to measure the level of students' higher-order thinking skills.

There are 25 valid questions and 15 invalid, according to the validity of the evaluation instrument. As in Syarif &; Syamsurizal (2019), questions are valid if there is a correlation between the scores on the question items and the scores on the question devices. Factors that affect the validity of questions are based on the suitability of the question instruments developed with good and correct evaluation instrument preparation procedures, disproportionate allocation of time for working on questions, the occurrence of cooperation when working on questions, the tendency of students to answer by trial and error or random (Gronlund, 1985). In this study, no revision of evaluation instruments has been carried out empirical validation tests. But researchers have made improvements according to the advice and direction of experts.

Next, using Kuder and Richardson's formula (KR-21), an evaluation of the reliability of evaluation instruments to measure higher-order thinking skills in acid-base materials was performed. By using this formula, the reliability value of the evaluation instrument is 0.815 which means that the question is in the very high reliability category. Sugiyono (2015) states that if the number of instrument reliability coefficient is greater than 0.70 (r1> 0.70), then the item is considered reliable. If the question is used repeatedly on the same student and the measurement results remain mostly the same, it is said to be reliable. Alternatively, it is said that the instrument criteria can be trusted if the test is used repeatedly and the measurement results remain consistent (Afrida et al, 2020).

The difficulty test of the evaluation instrument is carried out to determine the balance of difficulty of the developed questions. Good questions are those that are neither too challenging nor too simple. Based on the results of a study of 25 questions, there are 9 questions in the easy category, this means that almost all students can respond to these questions. Furthermore, there are 16 questions that fall into the medium category, meaning that some students can respond to questions correctly. Thus, the evaluation instrument developed has a good or sufficient level of difficulty. Based on research by Susanto et al (2015), the question bank book must be immediately updated with questions of medium difficulty. Exam questions can also be asked again next to assess learning outcomes. In the easy question instrument there are three possible follow-ups, namely: (1) Questions are discarded and will not be used in subsequent assessments of learning outcomes. (2) thoroughly re-examined and traced to identify the elements that caused the question item to be properly responded to by all test takers; After correction, the question item in question is issued again in the next test to determine whether the difficulty of the question item is increased or not during the initial example, (3) The use of easy question types in the loose selection test has the advantage that the majority of test takers will be considered successful in the selection test. In this condition, the availability of easy exam questions will give many students the opportunity to complete the test or selection exam given.

Discriminating power testing is carried out to distinguish students with above-average abilities from students with below-average abilities. According to Kusaeri (2014) the discriminating power on the question is said to be very good if it has a value range of 0.40-1.00, the discriminating power on the question is said to be good if it has a value range of 0.30-0.39, the discriminating power on the question is said to be sufficient if it has a value range of 0.20-0.29, and the discriminating power on the question is said to be bad if it has a value range of 0.00-0.19. Based on the results of the study, it is known that the distinguishing power of student evaluation instruments in the good category is 20%

or 5 questions, the sufficient category is 68% or 17 questions, and the bad category is 12% or 3 questions. This means that only 22 evaluation instrument questions have good distinguishing power in distinguishing high-ability students from low-ability students.

The next feasibility analysis of the evaluation instrument is the effectiveness of the distraction. The results of the distractor effectiveness analysis found that there were as many as 19 questions that were included in the criteria for the effectiveness of the distractor to be accepted, meaning that the deceiver could function properly. While there are 6 questions that fall into the category of effectiveness of the revised distractor, it means that the deceiver has not functioned properly. A distractor is an alternative option that distracts students from the correct answer on the test so that students are interested in choosing it. The more test takers choose a distraction, the better the distractor will function. Distractors work well if chosen by at least 5% of test takers (Arikunto, 2010). Based on research conducted by Amelia (2016), Distractors selected by less than 5% of test takers need to be revised.

Students' higher-order thinking skills are measured using students' scores on evaluation instruments on acid-base material. In line with Aisah's research (2020) that analysis of students' higher-order thinking skills can be obtained through working on the HOTS evaluation instrument. Based on research that has been conducted at SMA Negeri 2 Percut Sei Tuan, precisely in class XI A SMA, it was obtained that students' higher-order thinking skills on acid-base material, namely students have excellent HOTS with a range of 81-100 scores as much as 36% of the total students. Furthermore, in the range of 61-80 there are 31% of students who have high category abilities. In the sufficient category with a range of 41-60 there are 19% of students and the low category with a range of 21-40 there are 14%. While in the very low category 0 or no students with very low higher order thinking skills. According to the results of testing to students in class XI A SMA, the average student answers as many as 36% of questions with the level of cognitive analyzing (C4), 52% of the level of evaluating (C5), and 12% of the level of creating (C6). The average score of the students as a whole was obtained at 68.13 which means that the level of students' higher-order thinking skills falls into the good category.

One thinking skill, the ability to think higher, requires higher abilities such as analyzing, synthesizing, and judging in addition to memory. When a person learns new information, higher-order thinking skills emerge because the information is stored in memory and connected to other pieces of knowledge to complete tasks or find solutions to confusing situations. A person's higher-order thinking skills can be trained by getting used to solving complex problems. In this study, alternative solutions have been selected to help hone students' higher-order thinking skills by providing HOTS evaluation instruments. With this instrument, students are expected to be accustomed to solving complex questions that require students to analyze, evaluate, and synthesize based on existing problems.

Students' generic science skills support their capacity to complete the HOTS evaluation tool. There is generic competence because studying the ideas of science and solving problems basically follows the same process (according to the principle of the Nature Assessment Triangle). Students are encouraged to be able to think about science in their daily activities by having generic science skills, This is in accordance with the purpose of higher order thinking skills, which is to improve students' thinking skills at a higher level, especially those related to the capacity to think critically about the information they receive from various sources, think creatively about the problems they

face, and use their knowledge to make decisions in challenging circumstances (Saputra, 2016).

▪ CONCLUSION

Based on the results of research on the development of evaluation instruments to measure higher-order thinking skills in acid-base material in the following conclusions: An evaluation instrument has been developed to measure higher-order thinking skills in acid-base materials with a 4D development model, where this model consists of define, design, develop, and disseminate stages. Evaluation instruments to measure higher-order thinking skills in acid-base materials developed have met the eligibility criteria of good instruments including validity, reliability, level of difficulty, differentiation, and distraction. High-level thinking skills of grade XI A SMA Negeri 2 Percut Sei Tuan students on acid-base material are in the good category indicated by the average score of students which is 68.13.

▪ BIBLIOGRAPHY

Afrida., Sari, R. P., & Setianingsih, Y. (2020). Analysis of the Quality of Odd Semester Exam Question Items in Class V MI Civic Education Subjects. *Jurnal Keilmuan dan Kependidikan Dasar,* 12(02), 113-124.

Aisah,S., & Pahlevi, T. (2020). Development of Higher Order Thinking Skills (HOTS) Assessment Instruments in Class X OTP Correspondence Subjects at SMK Negeri 1 Jombang. *Jurnal Pendidikan Administrasi Perkantoran*, 8(1), 146-156.

Alfajri, A. R., Maizora, S., & Agustinsa, R. (2019). Practicality of Higher Order Thinking Questions to Produce Practical Questions for Class XI Students of MAN 1 Bengkulu City. *Jurnal Penelitian Pembelajaran Matematika Sekolah*, 3(2), 205-217.

Amelia, M. A. (2016). Analysis of High Order Thinking Skills (Hots) Mathematics Learning Outcomes Test Questions for Grade 5 Elementary School. *Jurnal Penelitian (Edisi Khusus PGSD)*, 20(2), 123-131.

Anderson, L. W., & Krathwohl, D. R. (2002). *Revision of Bloom's Taxonomy*. Jakarta: Rineka Cipta.

Anggreani, C., Permanasari, A., & Heliawati, L. (2022). Students' Scientific Literacy in Chemistry Learning through Collaborative Techniques as a Pillar of 21st-Century Skills. *Journal of Innovation in Educational and Cultural Research,* 3(3), 457-462.

Arikunto. (2010). *Research Procedure: A Practice Approach*. Jakarta: Rineka Cipta.

Brookhart, S.M. (2010*). Assess Higher Order Thingking Skills in Your Classroom*. Alexandria: ASCD.

Danggus, G. (2014). Improving Learning Outcomes of Polymer Material through the Application of the Numbered Heads Together Type Cooperative Learning Model to Class XII Science Students of SMAN 2 Pontianak. *Jurnal Penddikan Matematika dan IPA,* 5 (2), 9–20.

Drechsler, K., & Schmidt, H. (2005). Upper Secondary School Students' Understanding of Model's Used in Chemistry to Define Acids and Bases. *Science Education International*, 16(1), 39-53.

Fanani, M. Z. (2018). Higher Order Thinking Skill (HOTS) Question Development Strategy in the 2013 Curriculum. *Journal of Islamic Religious Educations,* (2)(1),

57-76.

Gronlund, N. E. (1985). *Measurencet and Evaluation in Teaching.* New York: Mc. Millan Publishing Co.

Kemendikbud. (2017). *Guidelines for Assessment by Educators and Education Units for Senior High Schools*. Jakarta: Direktorat Jenderal Pendidikan Dasar dan Menengah.

Khodijah., Marsani., & Murni. (2021). Correlation of Making HOTS Questions to Improving the Competence of Supervisors, Principals, Teachers, and Students. *Jurnal Ilmu Pendidikan (JIP),* 2(2), 80-86.

Kurniawati, F. N. A. (2022). Reviewing the Problem of Low Quality of Education in Indonesia and Solutions. *Academy of Education Journal*, 13(1), 1-13.

Kusaeri. (2014). *References and Assessment Techniques for Learning Processes and Outcomes in the 2013 Curriculum*. Yogyakarta: Ar-Ruzz Media.

Kusuma, M. D., Rosidin, U., Abdurrahman., & Suyatna, A. (2017). The Development of Higher Order Thinking Skill (Hots) Instrument Assessment in Physics Study. *Journal of Research and Method in Education*,7(1), 26-32.

Merta, I. W., Lestari, N., & Setiadi, D. (2019). Higher Order Thinking Skills (HOTS) Instrument Preparation Technique for Teachers of SMP Rayon 7 Mataram. *Jurnal Pendidikan dan Pengabdian Masyarakat*, 2(1), 1-10.

Nurkholik, M., & Yonata, B. (2020). Implementasi Model Pembelajaran Inkuiri Untuk Melatihkan High Order Thinking Skills Peserta Didik pada Materi Laju Reaksikelas XI IPA MAN 2 Gresik. *Unesa Journal of Chemical Education,* 9(1), 158-164.

OECD. (2019). PISA 2018. *PISA 2018 Result Combined Executive Summaries*. PISAOECD Publishing.

Salirawati., Permanasari, L., Purtadi, S., Nugraheni, A. R. E., & Dina. (2017). HOT (Higher Order Thinking) Question Development Training as an Increase in Teacher Pedagogic Competence. *Jurnal Relawan Indonesia*, 21(1), 14-25.

Sani, R. (2019). *HOT (Higher Order Thinking Skill) Based Learning*. Tangerang: Tira Smart.

Saputra, H. (2016). *Education Quality Development Towards the Global Era: Strengthening Learning Quality with the Application of HOTS (High Order Thinking Skills).* Bandung: SMILE's Publishing.

Saraswati, P. M. S., & Agustika, G. N. S. (2020). Higher Order Thinking Ability in Solving HOTS Math Problems. *Jurnal Ilmiah Sekolah Dasar*, 4(2), 257-269.

Stieff, M., & Wilensky, U. (2003). Connected Chemistry-Incorporating Interactive Simulations into the Chemistry Classroom. *Journal of Science Education and Technology*, 12(3), 285-302.

Sugiyono. (2015). *Metode Penelitian Kuantitatif Kualitatif dan R&D*. Bandung: Alfabeta.

Susanto, H., Rinaldi, A., & Novalia, N. (2015). Analysis of Validity of Reliability, Level of Difficulty and Differentiation in Class XII Social Studies Odd Semester Final Exam Question Items at SMA Negeri 12 Bandar Lampung Academic Year 2014/2015. *Jurnal Ilmu Pendidikan*, 6(2), 203-217.

Syarif, E. A., & Syamsurizal, S. (2019). Analyzed Quality of Senior High School Biology Olympiad Questions at West Sumatera, Riau, Jambi, and Bengkulu in 2018.*Bioeducation Journal*, 3(02), 142-150.

Thiagarajan, S; Semmel, D.S; & Semmel, M.I. (1974). *Instructional Development for Training Teachers of Exceptional Children: A Sourcebook.* Indiana: Indiana

University.

Yustika, A., Susatyo, E. B., & Nuswowati, M. (2014). Test Criteria for Chemistry Learning Outcomes Assessment Instrument, *Jurnal Inovasi Pendidikan Kimia*, 8(2), 1330-1339.