THE DEVELOPMENT OF HIGHER ORDER THINKING SKILLS (HOTS) TEST INSTRUMENT ON METABOLISM TOPIC FOR SENIOR HIGH SCHOOL LEVEL

Muhibbuddin[1*], Maria Ulfa[2], Andi Ulfa Tenri Pada[3], Hafnati Rahmatan[4], Hasanuddin[5]

[1]Dr., Universitas Syiah Kuala, Banda Aceh, Indonesia, muhib.bio@gmail.com
[2]Universitas Syiah Kuala, Banda Aceh, Indonesia, mariaulfa@gmail.com
[3]Dr., Universitas Syiah Kuala, Banda Aceh, Indonesia, andi_ulfa@unsyiah.ac.id
[4]Dr., Universitas Syiah Kuala, Banda Aceh, Indonesia, hafnati_rahmatan@unsyiah.ac.id
[5]Dr., Universitas Syiah Kuala, Banda Aceh, Indonesia, hasanuddin@fkip.unsyiah.ac.id
*Corresponding Author

Abstract

Research and development of the HOTS test instruments was carried out to measure students' learning outcomes in the HOTS learning process on Metabolism topic at senior high school level. This study used the research and development method of the Dick & Carey systems approach model which consists of nine stages, namely: 1) assess needs to help identify learning goals; 2) conduct instructional analysis and analyze learners and contexts; 3) write performance objectives; 4) develop assessment instruments; 5) develop instructional strategies; 6) develop and select instructional materials; 7) design and conduct formative evaluation; 8) revise instruction based on the formative evaluation; and 9) design and conduct summative evaluation. The data of the students' mastery on Metabolism topic was collected through a HOTS-based multiple-choice test related to Metabolism topic. A total of 610 research subjects were involved in this study. Content validity analysis were conducted using Aiken's V formula and empirical validity analysis were carried out using product moment Pearson correlation. The results reveal that based on content validation, 83 out of 100 test items are valid, proven by the Aiken's V value on the three indicators which is greater than 0.677. Based on empirical validation, in cycle I there are 60 valid test items as $r_{count} > r_{table}$ (0.12) and sig. 2 tailed value < α 0.05. In cycle II, there are 55 valid test items as $r_{count} > r_{table}$ (0.11) and sig. 2 tailed value < α 0.05. The 55 test items are in good quality category, thus they are ready to be implemented.

Keywords: Test Instrument, HOTS, Metabolism, content validity, empirical validity

1. INTRODUCTION

Learning can be defined as a process of behavior change that occurs in a person as the result of a learning experience. These changes can be observed at behavioral and cognitive (knowledge) aspects. Facing the challenges of the 21st century, Indonesian education is directed towards Higher Order Thinking Skills (HOTS)-

based learning. HOTS-based learning is expected to be able to create competitive and skillful outputs of 21st century skills, namely creativity and innovation, critical thinking and problem solving, collaboration, and communication. Consequently, a HOTS-based assessment design becomes an essential aspect in HOTS-based learning in order to meet this goal.

The implementation of the 2013 curriculum in Indonesia has directed Indonesian students towards a learning environment that support the development of higher order thinking skills. However, HOTS-based assessment is infrequently applied in measuring students' learning outcomes in HOTS learning. A study conducted by Utomo in 2018 revealed that Indonesian students had difficulty in solving science test items that involved analysis, evaluation, and creation, the three key aspects or the Operational Verbs (in Indonesian: Kata Kerja Operasional or KKO) of HOTS. This difficulty is caused by the infrequent exposure of HOTS assessment among Indonesian students.This result is in line with a research conducted by Tsaparlis in 2020 found out that some Indonesian students do not possess sufficient abilities to answer HOTS-based test items.

There are two main trends that encourage reformation in science learning, namely: (1) the belief that students need to develop HOTS; (2) students must possess deep understanding in the subjects they are learning, rather than just applying algorithms to solve problems. Higher Order Thinking (HOT) and Higher Order Thinking Skills (HOTS) should be contrasted with Lower Order Thinking (LOT) and Lower Order Thinking Skills (LOTS) in order to gain a clearer understanding of HOT and HOTS. Therefore, the measurement tools utilized in the assessment process must be in accordance with the learning objectives, meaningful and contributing to student achievement.

Up to now, there are still some limitations in the assessment implementation which leads to an off-target assessment. An initial survey of 52 high school biology teachers in Aceh using a questionnaire distributed online with the Google Form shows that the teachers's ability in developing test instruments that meet these criteria are still inadequate. 65.38% of the respondents admitted that they do not have the competence to prepare accurate test items according to the guidelines and 69.23% of them claimed to be more comfortable to reuse test items that were already available to assess student learning outcomes.

The results of the survey is also supported by the study results of item analysis conducted by the Directorate of High School Development for USBN Assistance for the 2018/2019 academic year. This study was conducted on 26 subject matters taught in 136 premier high schools spread across 34 Provinces in Indonesia. The results reveal that from 1,779 items analyzed, most of which were at Level-1 and Level-2. Furthermore, only 27 out of 136 premier high schools constructed 20% of USBN (The National School-Based Examination) test items based on HOTS criteria, 84 schools constructed below 20%, and 25 schools stated that they did not know whether the test items constructed met the HOTS criteria or not. These facts do not meet the Curriculum 2013 assessment standard which aims to improve the implementation of HOTS assessment models (Isbandiyah & Sanusi, 2019).

The results of the Trends in International Mathematics and Science Study (TIMMS) in 2015 for grade IV of elementary school levels showed that Indonesia got an average score of 397 and was ranked the bottom four of the 43 countries participating in TIMMS (Source: TIMMS 2015 International Database). Approximately, 75% of the items tested in TIMSS have been taught in grade IV at elementary schools. This percentage is higher than South Korea which is only 68%, but the depth of understanding towards the test items is still lacking. Ironically, Indonesia is the longest among other countries in terms of the length of school hours and duration of mathematics subject hours at elementary school level, but the quality of learning outcomes still needs to be improved. In addition, the *Programme for International Student Assessment* (PISA) in 2015 shows that Indonesia received an average score of 403 for science (ranked the bottom three), 397 for reading (last rank), and 386 for mathematics (ranked the bottom two) from 72 countries. In 2018, the average PISA score of Indonesian students decreased. Indonesia received an average score of 396 for science, 371 for reading, and 379 for mathematics which was ranked in the bottom six of the 79 countries participating the program (Setiawati, et. al, 2018).

The results of student achievement measurement of the National Examination are along with students' achievements in PISA and TIMSS which show that students are still weak in higher order thinking skills such as reasoning, analyzing, and evaluation. This statement is confirmed by the data of the average National Examination (UN) score distribution at high school levels in both public and private schools in the last four years. It shows that the average achievement of national examination results is 57.29 (2016); 53.47 (in 2017); 51.76 (in 2018); and 53.00 (in 2019) within the range of 2.5-100 point scale. The average achievement of the 2019

National Examination results in Aceh Province was 43.03, which was 9.27 points lower than the average National Examination result of 52.30 (Pusmenjar Kemdikbud, 2019). According to the Ministry of Education and Culture, an acceptable standard of National Examination score is 70-85. Thus there is a huge gap between the average score achieved by the students and the acceptable standard of National Examination (UN) scores expected by the Ministry of Education and Culture. In addition, there is a great disparity between the students' highest score and the lowest one. This indicates that the quality of education in Indonesia, especially in Aceh Province, still needs to be improved and the design of the assessment process should be enhanced.

Several studies on assessment especially on the development of HOTS-based instruments have been carried out (Rintayati, P, 2021; Su, 2020; Hidayat, 2020; Luo, 2020; Rahayu, 2020; Subia, G.S, 2020; Risna, 2020; Kurniawan, 2019 ; Istiyono, E, 2019; Semilarski, 2019; Wahono, 2019; Utomo, 2018; Hamdi, 2018; Kusuma, 2017; Ahmad, 2017; Kurbanoglu, 2017; Barz, 2017; Saat, 2016; Samritin, 2016; Hanum, 2015 ; Park, 2014; Muhammad, 2014; Lemons, 2013; Koksal, 2010). However, their study only focus on the development of diagnostic instruments, the introduction and development of the HOTS instruments in general to measure students' scientific reasoning skills to achieve the target scores of PISA and TIMSS international standards. The study on development of HOTS-based test instruments that focus on certain cognitive competencies, especially on the basic competence of Metabolism is still limited. There is an urgent demand for providing guidelines in preparing HOTS test items that can be implemented by teachers in classroom learning assessment activities for the basic competence.

Based on data analysis of the National Examination scores of high school level on Biology subject conducted by the Center for Educational Assessment of the Ministry of Education and Culture, the average score in four topics of Biology subject tested in the National Examination, namely (1) biodiversity and ecology; (2) the structure and function of living things; (3) biomolecular and biotechnology; and (4) genetics and evolution, the biomolecular and biotechnology topic get the lowest average score, namely; 47.25 (in 2017); 48.57 (in 2018); and decreased to 39.70 (in 2019). Unsurprisingly, amidst the development of science and technology in the 4.0 era, these competencies are expected to increase even more.

The findings ware also supported by the results of an initial survey conducted on Biology teachers at the high school level in Aceh. 69.23% of teachers stated that Metabolism was a difficult topic for both teachers and students, 61.54% admitted that the students' learning outcomes on Metabolism was always low. 69.23% claimed that the students' learning outcomes is due to inadequate learning process and 75% of teachers also agree it is caused by inaccurate test instrument. Responding to this phenomenon, 100% of teachers admitted that they need guidelines for the preparation of valid test instruments.

In order to meet these needs, a valid HOTS instrument for Metabolism topic must be accessible. Therefore, a research on the development of Higher Order Thinking Skills (HOTS)-test instrument on Metabolism topc for high school level needs to be conducted. This research is expected to be able to provide a valid measurement tool guidelines in Metabolism Basic Competence (KD) and to improve the higher-order thinking skills of students in Aceh in particular and students in Indonesia in general.

2. METHODS

This study used the research and development method of the Dick & Carey systems approach model which consists of nine stages, namely: 1) assess needs to help identify learning goals; 2) conduct instructional analysis and analyze learners and contexts; 3) write performance objectives; 4) develop assessment instruments; 5) develop instructional strategies; 6) develop and select instructional materials; 7) design and conduct formative evaluation; 8) revise instruction based on the formative evaluation; and 9) design and conduct summative evaluation (optional) (Dick et al., 2001). The design of the Dick & Carey systems approach model phases is illustrated in Figure -1.

The total subjects involved in this study were 610 students which were divided into 3 cycles: 250 subjects in Cycle 1 (Formative Test), 300 subjects in Cycle II (Summative Test) and 60 subjects in the implementation phase, consisting of an experimental class of 30 students and a control class of 30 students. Ths study used probability sampling technique. The research was conducted in eight months from February-September 2021 in seven high schools in Aceh Province, Indonesia. The parameters measured were HOTS-based learning outcomes on Metabolism topic obtained through pretest and posttest, totaling 100 test items.

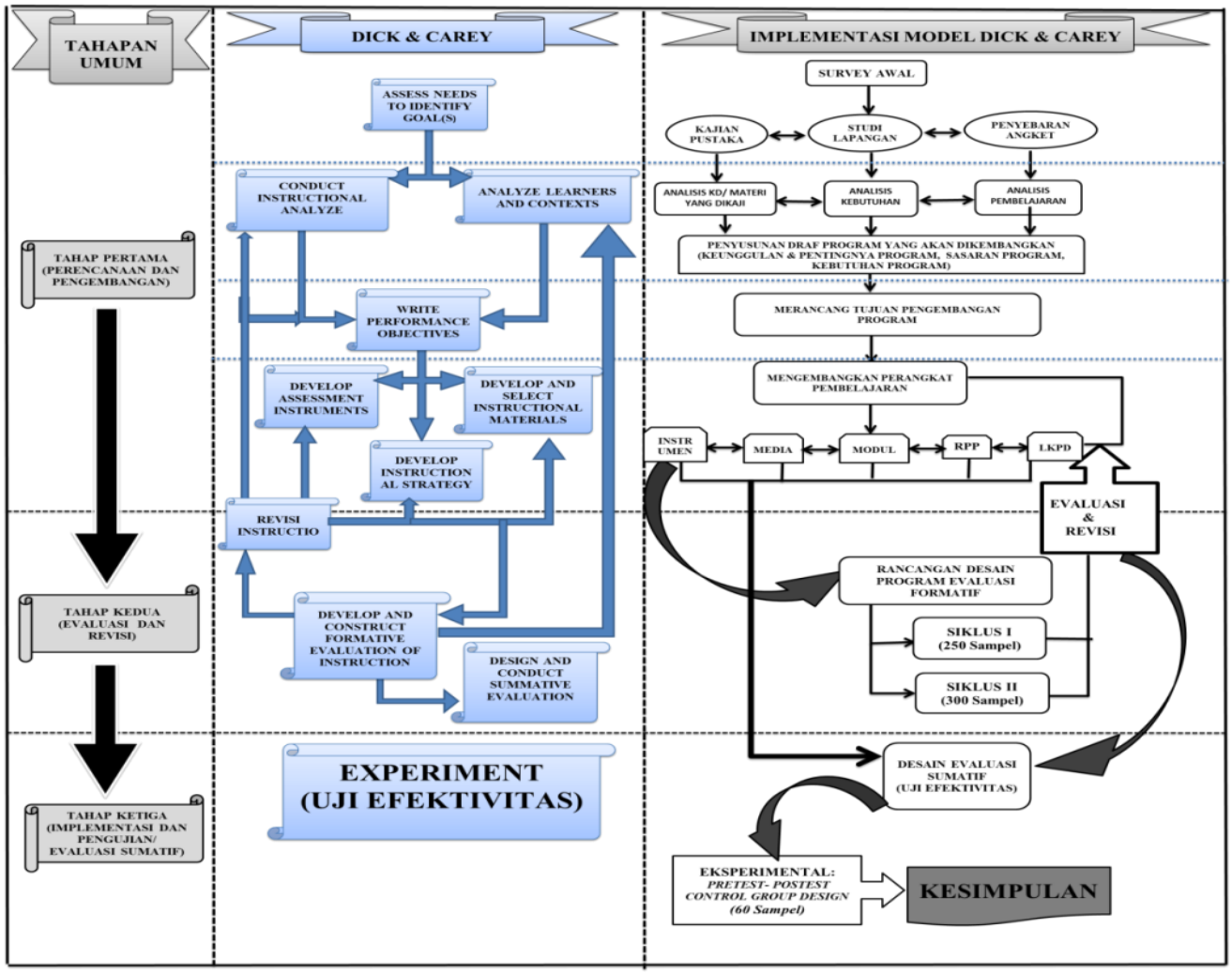

Figure-1. The design of the Dick & Carey systems approach model phases

3. DATA AND ANALYSIS

Content validity analysis of the HOTS-based test instrument was conducted using Aiken's V coefficient which was calculated by the formula V = S/[n(c — 1)] (Aiken, 1985). A good assessment formation will produce consistent scores and positive responses from validators with an average Aiken's V assessment score greater than 0.677 (Azwar, 2014). empirical validity analysis of cycle I and Cycle II were carried out using product moment Pearson correlation which correlate each test item score with the total score obtained from all respondents' answers.

4. FINDINGS AND DISCUSSION

4.1 Content Validation of HOTS Test Instrument

The instrument that has been developed consists of 100 test items in the form of multiple choice to evaluate the students' learning outcomes on Metabolism. The instrument need to be first evaluated by the experts before it can be administered. The content validity was carried out by distributing validation sheets to 3 (three) validators who are the experts of the field being studied. In the early stages of instrument development, the content validity analysis was aimed to reduce variations of potential errors in the instrument development and increase the possibility of obtaining a construct validity index in future research (Muliana, 2020). There are three indicators evaluated through the validation sheets, namely the suitability of the question indicator formulation (indicator 1), the suitability of the indicators with Bloom's taxonomy level (indicator 2), and the suitability of the question diction

with the question stimulus (indicator 3). Those indicators were evaluated using a 5-point scale. The content validity results of the three indicators are presented in Figure-2.

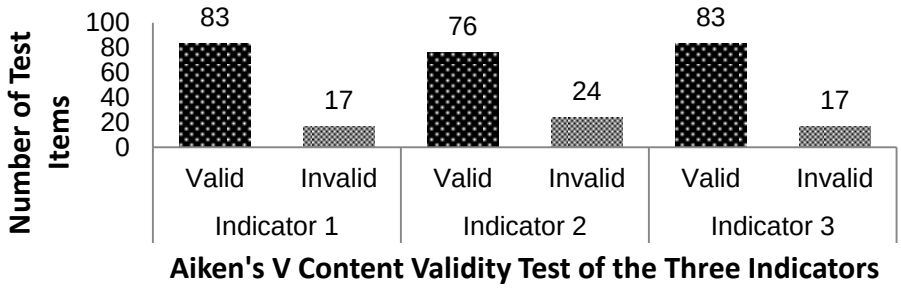

Figure-2. The Recapitulation of Content Validity Test Results with the Aiken's V-Formula for the three Indicators

The content validity result for indicator 1 shows that 83 test items are valid as the Aiken's V score > 0,677. However, 17 test items are invalid because the Aiken's V score ≤ 0,677. Those invalid test items are items number 12,14, 15, 17, 30, 32, 41, 42, 53, 68, 70, 80, 82, 83, 84, 95, and 100.

The content validity result for indicator 2 shows that 76 test items are valid as the Aiken's V score > 0,677. Meanwhile, 24 test items are invalid because the Aiken's V score ≤ 0,677, namely item number 1, 2, 3,12, 13, 14, 15, 17, 28, 30, 32, 41, 42, 50, 53, 68, 70, 78, 80, 82, 83, 84, 95, and 100. Muliana (2020) stated that the more raters involved, the smaller the index of the criteria specified and vice versa, the less raters involved, the greater the index of validity of the criteria specified. However, Azwar said that a coefficient between 0.64 to 1 can be considered to have sufficient content validity.

The content validity result for indicator 3 shows that 83 test items are valid as the Aiken's V score > 0,677. Notwithstanding, 17 test items are invalid as the Aiken's V score ≤ 0,677, Those are the item number 12,14, 15, 17, 30, 32, 41, 42, 53, 68, 70, 80, 82, 83, 84, 95, and 100.

Based on the validation results of those three indicators, 7 items are invalid only on the indicator 1, namely the item number 1, 2, 3, 13, 28, 50, and 78. Thus, the test items are still used after revision. However, 17 items are invalid on the three indicators because the Aiken's V score on the three indicators is below 0.677, consequently, all these test items are discarded. The items eliminated are number 12,14, 15, 17, 30, 32, 41, 42, 53, 68, 70, 80, 82, 83, 84, 95, and 100.

4.2 Empirical Validation of HOTS Test Instrument

4.2.1 Empirical Validation of Cycle I (Formative Evaluation)

In the field test phase of cycle I (formative evaluation), a total of 83 test items that have been expertly validated and revised were tested on 250 subjects from five (5) schools in Aceh Besar and Banda Aceh. After the distribution of the field test data of Cycle I was normal, the validity test was carried out using SPSS 22 for windows program. A test item is said to be valid if it can accurately measures what it is supposed to measure.

At this stage, the product moment Pearson correlation validity test was used. Each test item score was correlated with the total score obtained from all respondents' answers. The empirical validity results of Cycle I are presented in Figure-3.

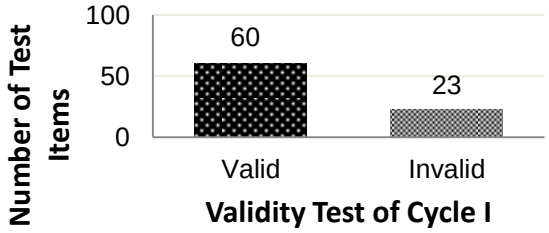

Figure-3. The empirical validity results of Cycle I

The validity of the test items is calculated by comparing the correlation coefficient (pearson correlation) scores of each test item with r_{table} or by comparing the value of sig. 2 tailed with a significant level α = 0.05. A test item is said to be valid if $r_{count} > r_{table}$ with a significant value α = 0.05 or sig. 2 tailed value α < 0.05. Based on the results of data analysis in Cycle I, the Pearson correlation (r_{count}) score and sig. 2 tailed of 83 test items tested on 250 samples were obtained The r_{table} value with N = 250 at the significant level α = 0.05 is 0.12. Therefore, there are 60 valid test items because $r_{count} > r_{table}$ (0.12) and the value of sig. 2 tailed < α 0.05. Yet, there are 23 invalid test items because r $r_{count} < r_{table}$ (0.12) and sig. 2 tailed value > α 0.05, namely test items number 2, 3, 19, 25, 38, 45, 48, 52, 54, 57, 59, 60, 63, 64, 66, 67, 69, 72, 74, 76, 77, 79, and 80.

The correlation coefficient of the test items were categorized according to the interpretation of the correlation coefficient of Guilford. This interpretation was carried out to determine the high, moderate or low relationship between the two factors. The standard rules provided by Guildford (1973) were adopted to interpret the strength of the relationship. Table-1 presents a summary of the standard rules for the interpretation of the correlation coefficient (r) based on Guildford (1973).

Table-1 Guildford's (1973) Rule of Thumb for Interpretation of Correlation Coefficient (r)

r	**Interpretation**
<0,2	Negligible positive/negative correlation
0,2-0,4	Low positive/negative correlation
0,4-0,7	Moderate positive/negative correlation
0,7-0,9	High positive/negative correlation
> 0,9	Veri High positive/negative correlation

Based on *Pearson correlation* value, the distribution of the HOTS test items quality in cycle I is presented in Figure-4.

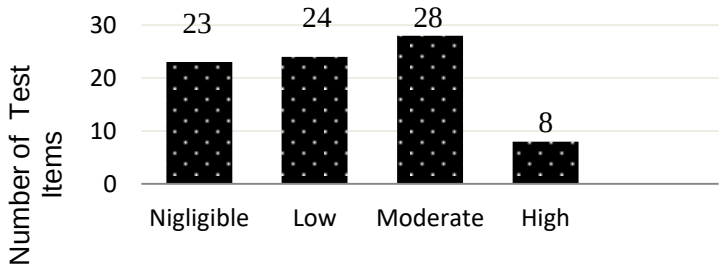

Figure-4. The distribution of the HOTS test items quality in Cycle I

Based on the Pearson correlation score above with the interpretation of the Guilford correlation coefficient, the test items with high correlation (8 items), moderate correlation (28 items), and low correlation (24 items) criteria were declared valid with some revisions, while the test items with very low/neglected correlation criteria (23 items) were declared invalid and no longer used in the next development stage. In other words, based on the validity test, from the 83 test items there were 23 test items that were eliminated and only 60 test items would be re-tested in cycle II after some revisions. The revisions made to the 60 test items included adjusting the cognitive level, adjusting the question indicators, improving the stimulus questions, and improving the narrative of the questions so that they are not ambiguous and confusing for the students.

4.2.2 Empirical Validation of Cycle II (Summative Evaluation)

After analyzing the test item validity of the Cycle I field test result (formative evaluation), 60 valid test items were produced after some revisions were made. In the field test phase of cycle II (summative evaluation), as many as 60 test items were tested on 300 samples from 7 senior high schools in Aceh province.

Subsequent of the distribution of the field test data of Cycle II was normal, the validity test was carried out using SPSS 22 for windows program. A test item is said to be valid if it can accurately measures what it is supposed to measure. At this stage, the product moment Pearson correlation validity test was used. Each test item score was

correlated with the total score obtained from all respondents' answers. The empirical validity results of Cycle II are presented in Figure-5.

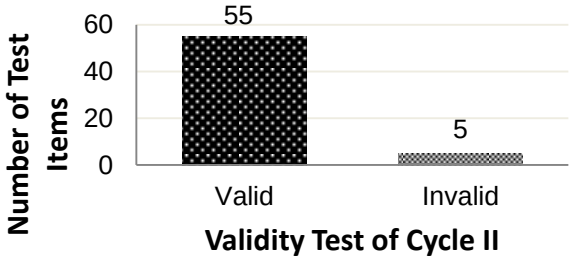

Figure-5. The empirical validity results of Cycle II

The validity of the test items is calculated by comparing the correlation coefficient (pearson correlation) scores of each test item with r_{table} or by comparing the value of sig. 2 tailed with a significant level α = 0.05. A test item is said to be valid if $r_{count} > r_{table}$ with a significant value α = 0.05 or sig. 2 tailed value α < 0.05. Based on the results of data analysis in Cycle II, the Pearson correlation (r_{count}) score and sig. 2 tailed of 60 test items tested on 300 samples were obtained The r_{table} value with N = 300 at the significant level α = 0.05 is 0.11. As the result, there are 55 valid test items because $r_{count} > r_{table}$ (0.11) and the value of sig. 2 tailed < α 0.05. Yet, there are 5 invalid test items because r $r_{count} < r_{table}$ (0.11) and sig. 2 tailed value > α 0.05, namely test items number 28, 44, 54, 55, and 56.

Based on the Pearson correlation score, The distribution of the HOTS test items quality in cycle II is presented in Figure-6. The correlation coefficient of the test items were categorized according to the interpretation of the correlation coefficient of Guilford. This interpretation was carried out to determine the high, moderate or low relationship between the two factors. The standard rules provided by Guildford (1973) were adopted to interpret the strength of the relationship.

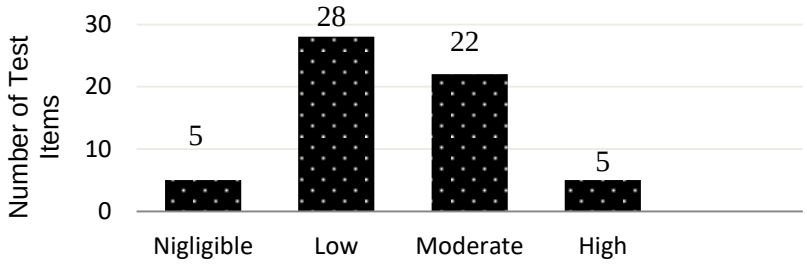

Figure-6. The distribution of the HOTS test items quality in Cycle II

Based on the Pearson correlation score above, the distribution of the HOTS test items quality in Cycle II are the test items with high correlation (5 items), moderate correlation (22 items), and low correlation (28 items) criteria that were declared valid with some revisions. However, the test items with very low/neglected correlation criteria (5 items) were declared invalid and no longer used in the next development stage. In other words, based on the validity test of those 60 test items, 5 test items were eliminated and only 55 test items would be used in the implementation stage after some revisions.

5. CONCLUSION

Based on the research and development results of the HOTS test instrument with the Dick & Carey systems approach model, 83 test items out of 100 test items that were developed based on content validation are valid because the Aiken's V value on the three indicators was greater than 0.677. Based on the empirical validation, there are 60 valid test items in Cycle I because $r_{count} > r_{table}$ (0.12) and sig. 2 tailed < α 0.05. Then, there are 55 valid test items as $r_{count} < r_{table}$ (0.11) rhit > rtable (0.11) and sig. 2 tailed < α 0.05. The 55 test items are in good quality so they were ready to be implemented.

REFERENCE LIST

Ahmad, S.; Prahmana, R.C.I.; Kenedi, A.K.; Helsa, Y.; Arianil, Y.; & Zainil, M. (2017). The Instruments of Higher Order Thinking Skills. IOP Conf. Series: *Journal of Physics,* 9(4): 20-36.

Azwar, S. (2014). *Reliabilitas dan Validitas*. Yogyakarta: Pustaka Pelajar.

Barz, D.L. & Cadariu, A.A. (2017). Development Of A Skillsbased Instrument To Measure Scientific Reasoning In Medicine Across Different Levels Of Expertise. *Journal of Baltic Science Education*, 16(3): 289-299.

Dick, W; Carey, L, & Carey, J.O. (2001). *The Systematic Design of Instructio*. New York: Longman.

Guilford, J. P. (1973). Fundamental statistics in psychology and education. New York, NY: McGraw-Hill.

Hamdi, S.; Suganda, I.A. & Hayati, N. (2018). Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics. *REiD (Research and Evaluation in Education)*, 4(2): 126-135.

Hanum, E.; Yusrizal & Muhibbuddin. (2015). Analisis Tingkat Kesukaran Dan Reliabilitas Dalam Pengembangan Item Tes Keterampilan Proses Sains Biologi*. Jurnal EduBio Tropika,* 3(2): 51-97.

Hidayat, D.M.F.; Adisaputera, A. & Pamuniati, I. (2020). Development of HOTS (High Order Thinking Skill) Based News Text Assessment Instrument for 8th Grade Students in SMP Muhammadiyah 7 Medan. *Budapest International Research and Critics in Linguistics and Education (BirLE) Journal.* 3(2): 1123-1136.

Isbandiyah, S. & Sanusi, A. (2019). *Modul Penyusunan Soal Keterampilan Berpikir Tingkat Tinggi Mata Pelajaran Biologi*. Jakarta: Direktorat Pembinaan Sekolah Menengah Atas Direktorat Jenderal Pendidikan Dasar Dan Menengah Kementerian Pendidikan Dan Kebudayaan

Istyono, E.; Dwandaru, W.S.B.; Setiawan, R. & Megawati, I. (2019). Developing of Computerized Adaptive Testing to Measure Physics Higher Order Thinking Skills of Senior High School Students and its Feasibility of Use. *European Journal of Educational Research*, 9(1): 91-101.

Koksal, M.S. & Cakiroglu, J. (2010). Development Of Nature Of Science Scala (NSS) For Advenced Science Student. *Journal Of Baltic Science Education,* 9(2): 87-98.

Kurniawan, R.Y. & Lestari, D. (2019). The Development Assessment Instruments of Higher Order Thinking Skills on Economic Subject. *Dinamika Pendidikan*. 14(1): 102-115.

Kurbanoğlu, N.I. & Takunyaci, M. (2017). Development And Evaluation Of An Instrument Measuring Anxiety Toward Physics Laboratory Classes Among University Students. *Journal of Baltic Science Education*, 16(4): 592- 598.

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

Lemons, P.P. & Lemons, J.D. (2013). Questions for Assessing Higher-Order Cognitive Skills: It's Not Just Bloom's. *CBE—Life Sciences Education, Spring*, 12(2): 47–58.

Lou, M.; Wang, Z.; Sun, D.; Wan, Z.H. & Zhu, L. (2020). Evaluating Scientific Reasoning Ability: The Design And Validation Of An Assessment With A Focus On Reasoning And The Use Of Evidenc. *Journal Of Baltic Science Education,* 19(2):261-275.

Muhammad, N.; Djufri & Muhibbuddin. (2014). Penerapan Model Concept Attainment Terhadap Hasil Belajar Siswa Pada Materi Metabolisme. *Jurnal Biologi Edukasi,* 6(1): 9-15

Muliana, M.; Pada, AUT.; Nurmaliah,C. (2020). Content Validity Of Conation Assessment. *Journal Od Physics: Conference Series,* 14(6): 1-6.

Park, J.; Park, Y.S.; Kim, Y. & Joeng, J.S. (2014). The Development Of The Korean Teaching Observation Protocol (Ktop) For Improving Science Teaching And Learning. *Journal Of Baltic Science Education,* 13(2): 259-275.

Pusmenjar. (2018). *Laporan Hasil Ujian Nasional SMA/MA Tahun Pelajaran 2015-2019*. Jakarta: Pusat Penilaian Pendidikan Kemendikbud. Tersedia di: *https://pusmenjar.kemdikbud. go.id/hasil-un,* (Diakses Tanggal 23 Januari 2022).

Rahayu, P.W.; Hasanah, U. & Wiliandri, R. (2020). Developing a Higher Order Thinking Skill-Oriented and Metacognitive-Based Assessment for Vocational School Students. *Jurnal Pendidikan Bisnis dan Manajemen*, 6(1): 10-23.

Rintayati, P.; Lukitasari, H. & Syawaludin, A.(2021). Development of Two-Tier Multiple Choice Test to Assess Indonesian Elementary Students' Higher-Order Thinking Skills. *International Journal of Instruction*, 14(1): 555-566.

Risna; Hasan, M. & Supriatno. (2020). Implementation of guided inquiry learning oriented to green chemistry to enhance students' higher-order thinking skills. IOP Conf. Series: *Journal of Physics*: Conf. 14(6): 1-15.

Saat, R.M.; Fadzil, H.M.; Aziz, N.A.A.; Haron, K.; Rashid, K.A.; Shamsuar, N.R. (2016). Development Of An Online Three-Tier Diagnostic Test To Assess Pre-University Students' Understanding Of Cellular Respiration. *Journal of Baltic Science Education,* 15(4): 532-546.

Samilarski, H.; Laius, A. & Rannikmae, M. (2019). Development Of Estonian Upper Secondary School Students' Biological Conceptual Understanding And Competences. *Journal Of Baltic Science Education,* 18(9): 955-970.

Samritin & Suryanto. (2016). Developing An Assessment Instrument Of Junior High School Students' Higher Order Thinking Skills In Mathematics. *Research and Evaluation in Education (REID).* 2(1): 92-107.

Setiawati, W.; Asmira, O.; Ariyana, Y.; Bestary, R. & Pudjiastuti, A. (2018). *Buku Penilaian Berorientasi pada Keterampilan Berpikir Tingkat Tinggi-Program Peningkatan Kompetensi Pembelajaran Berbasis Zonasi.* Jakarta: Direktorat Jenderal Guru dan Tenaga Kependidikan Kementerian Pendidikan dan Kebudayaan.

Su, K.D. (2020). Enhancing Students' Highorder Cognitive Skills For Hierarchical Designs In Micro And Symbolic Particulate Nature Of Matter. *Journal of Baltic Science Education,* 19(5): 842-854.

Subia, G.S.; Marcos, M.C.; Pascual, L.E.; Tomas, A.V. & Liangco,M.M. (2020). Cognitive Levels as Measure of Higher-Order Thinking Skills in Senior High School Mathematics of Science, Technology, Engineering and Mathematics (STEM) Graduates. *Technology Reports of Kansai University*. 62(3): 261-268.

Tsaparlis, G. (2020). Higher And Lower-Order Thinking Skills: The Case Of Chemistry Revisited. *Journal of Baltic Science Education,* 19(3): 467-483.

Utomo, A.P.; Narulita, E. & Shimizu, K. (2018) Diversification Of Reasoning Science Test Items Of TIMSS Grade 8 Based On Higher Order Thinking Skill; A Case Study Of Indonesian Students. *Journal of Baltic Science Education,* 17(1): 152-161.

Wahono, B. & Chang, c.y. (2019). Development and Validation Of A Survey Instrument (Aka) Towards Attitude, Knowledge And Application of Stem. *Journal of Baltic Science Education*, 18(1): 63-76.