Advances in Social Science, Education and Humanities Research, volume 566
Proceedings of the 5th Asian Education Symposium 2020 (AES 2020)

Development of Classroom Assessment Tests Based on Higher Order Thinking Skills (HOTS)

Ni Ketut Widiartini*, Putu Putri Dena Laksmi
Graduate Program of Research and Educational Evaluation Department
Universitas Pendidikan Ganesha
Bali, Indonesia
*ketut.widiartini@pasca.undiksha.ac.id

***Abstract*—This study aims to describe the steps for the development and quality of the Classroom Based Assessment course test based on Higher Order Thinking Skills (HOTS). This research is a development research using a modified Borg & Gall model which consists of 4 main stages, namely needs analysis, initial product design, validation and evaluation, and final product design. Product testing to determine the quality of HOTS questions developed was carried out on 30 postgraduate students at PPs. Undiksha. The results showed, (1) the stages of test development included information gathering, grid development, preliminary design, expert validation, limited trials, initial product revisions, large group trials, final revisions, and final test design, and (2) the quality of the tests developed was classified as good with very high content validity, both in terms of material, HOTS criteria, and question construction (CVR 0.98). Furthermore, the quality of the empirical validity was good with 95% of the test items developed were valid and the reliability of the test kits was high (0.64). Meanwhile, the overall difficulty level of the test kits was classified as moderate (P 0.52) with 12% of the questions having an easy difficulty level, 86% of the questions having a moderate difficulty level, and 2% of the questions having a high difficulty level. Judging from the distinguishing power, this test kit has good differentiation power and as many as 77% of the distractor test items are functioning properly.**

Keywords—higher order thinking skills, class-based assessment, test development

I. INTRODUCTION

Implementing learning activities that can grow the competence of students and support for an assessment system that is able to measure the achievement of learning implementation in the 21st century is needed, namely an assessment system that adopts international standard assessment models that are able to measure the competence of students. The assessment of learning outcomes that is expected in addition to being able to measure competence, is also expected to help students to improve their higher order thinking skills (HOTS) because high-order thinking can encourage students to think broadly and deeply, so that the assessment instrument which is used in the assessment system must be able to measure and foster the ability to think critically, creatively, innovatively, and solve problems, the ability to adapt to the environment and technology, the ability to make decisions, as well as strong and positive characters in students.

According to Langreh, to train critical thinking students must be encouraged to answer questions related to, (1) determining the consequences of a decision or event, (2) identifying the assumptions used in a statement, (3) formulating problem points, (4) finding bias based on different points of view, (5) revealing the causes of an event, and (6) selecting factors that support a decision. Likewise, in carrying out the assessment of learning outcomes, in addition to fulfilling the basic principles of assessment, it must also meet the demands of 21st century skills, which must be able to measure students' mastery of the quality of character, competence, and literacy mastery, and be able to develop level thinking high processes [1].

In developing a good instrument to measure critical thinking skills, Abidin [2] suggests the preparation of an assessment instrument using an interpretation assessment approach. Assessment with an interpretive approach is an assessment or assignment that requires students to use reading material, graphics, tables, pictures, or other materials to answer questions. Therefore, if it is related to Bloom's taxonomy which is the general basis for measuring the success of learning, the appropriate assessment instrument to measure 21st century competence must be oriented towards higher order thinking skills and abilities (analysis, evaluation, and creation).

Higher order thinking skills are not the ability to remember, know, or repeat. Higher order thinking skills include problem-solving skills, critical thinking skills, creative thinking, argumentation skills, and decision-making skills. HOT test or questions do not just measure the ability to remember, restate, or refer without doing any processing. HOTS questions in the context of the assessment measure the ability to transfer one concept to another, process and apply information, find links from different types of information, use information to solve problems, and critically analyse ideas and information. Even

so, HOTS questions do not mean more difficult questions than recall questions [1].

The importance of HOTS-based assessment instruments is evident from the results of research by Lewy et al. [3] that by gradually giving a prototype test that was developed to measure higher order thinking skills, there was an increase in students' thinking skills. This means that HOTS-based tests can improve higher-order thinking skills. However, most teachers in senior secondary schools ignore and have not been able to compile these assessment instruments. This is evidenced by the fact that there are still many teachers who compile tests consisting of questions that come from textbooks, where after being analysed it turns out that the questions are only at the low order thinking skills level (C1-C3). Whereas in the 2012 BAN assessment standards, ideally 80% of questions cover questions on the criteria for high order thinking skills (C4-C6). The results of the PISA analysis also show that almost all Indonesian students in master lessons up to level 3, while many other countries have mastered lessons up to level 4, 5, even 6 [4]. Likewise, the results of the Supervision and Development of Post-Learning Outcomes Evaluation (EHB) for SMA that have been implemented by the Directorate of Senior High School Development, most SMA teachers in compiling items tended to only measure low-level thinking skills and the questions were made not contextually. The questions compiled by the teacher generally measure recall skills (recall / C1). When viewed from the context, most of them use the context in the classroom and are very theoretical, and rarely use contexts outside the classroom (contextual), so that they do not show the relationship between the knowledge gained in learning and real situations in everyday life. Meanwhile, the results of the study Stiadi [5] shows that many teachers do not yet understand about question grids and their uses, how to analyse instruments, and make scoring guidelines.

Based on the description above, a study is needed to develop an assessment instrument that can measure students 'abilities in higher-order thinking, increase creativity, and build students' independence in solving problems. Therefore, this study aims to describe the development steps and quality of tests based on higher order thinking skills.

II. Research Methods

This research is a research development (R&D) using the Borg & Gall model. The development procedure carried out based on the theory of Borg and Gall [6] develops mini learning (mini courses) through 8 steps 1) Conducting preliminary research, 2) Planning, 3) develop the type / form of the initial product, 4) Conducting expert judgment, 5) Make revisions to the main product, 6) Conduct limited trials on each of the subjects of each competency and each character item, 7) field trials with subjects each school, 8) Conducting a final revision to obtain a ready-to-use instrument. In this research, the Borg & Gall instrument development model is modified according to the time and scope of the research without reducing the meaning of the borg and gall theory. There are four stages, namely, (1) the preparation stage, (2) the initial product design stage, (3) the validation and evaluation stage, and (4) the final product design stage. The research population was all PPs Undiksha students, which took class-based assessment courses, consisted of 30 students and research was carried out on all these students. The research data were collected using a check list and HOTS test instruments. The data is then analysed using qualitative data analysis techniques (to examine the validity of the content) and quantitative data analysis techniques (to analyse empirical validity, reliability, difficulty level, differentiation power, and deceiving effectiveness).

III. Research Results and Discussion

This research was conducted by developing a test kit based on higher order thinking skills in the class-based assessment course. Then, to determine the quality of the test, a trial was carried out in a lecture situation. The research results and discussion are as follows.

A. Preparation Phase

The activity at the preparation stage is to analyse basic competencies (KD) together with a team of lecturers from the Education Research and Evaluation study program to determine basic competencies which are categorized as high order thinking skill. In table 1 below is the selected KD.

TABLE I. Mapping of Basic Competencies for HOTS Categories

Basic Competencies	**Cognitive Process**						**Knowledge**				**Note**
	LOTS			***HOTS***			***Dimensions***				
	C1	***C2***	***C3***	***C4***	***C5***	***C6***	***K1***	***K2***	***K3***	***K4***	
3.1 Analyzed Classical test theory				√				√	√		C4,K2 C4,K3
3.2 Analyzed Item Response Theory				√				√		√	C4,K2 C4,K4
3.3 Analyzed the HOTS concept				√						√	C4,K4
3.4 Analyzed Models Item Response				√				√			C4,K2
3.5 Determine HOTs Instrument in IRT				√					√		C4,K3
Total		0		5							
Percentage		-		100%							

B. Initial Product Design Stage

Based on the results of the formulation of competency attainment indicators, test items based on higher-order thinking skills were compiled. The test developed is a multiple-choice test with 5 (five) alternative answers. There are 5 basic competency achievements that can be developed into basic competency indicators that will be developed with the HOTs test model, namely, 1) Analysed Classical test theory; 2) Analysed Item Response Theory; 3) Analysed the HOTS concept; 4) Analysed Models Item Response ; 5) Determine HOTs Instrument in IRT.

Each basic competency is redeveloped into several indicators of competency achievement that will be achieved which later can provide a reference for the preparation of test items. In the first basic competency, Analysed Classical test theory keep 5 performance indicators are developed as follows.

1. Analysed classical test concepts and methods, developed two test items with dimensions of knowledge on evaluating and cognitive processes that contain the conceptual results of teste knowledge on test instrument items 1 and 2.
2. Analysed classic test requirements, developed two test items with the dimensions of knowledge in analysing and cognitive processes that contain procedural knowledge of the testees on test instrument items 3 and 5.
3. Analysed the concept of calculating validity and reliability, developed two test items with the dimensions of knowledge in analysing and cognitive processes that contain the conceptual results of testee knowledge on test instrument items 4 and 6.
4. Analyse the concept of item difficulty index, and the power of difference two test items were developed with the dimensions of knowledge in evaluating and cognitive processes containing the conceptual results of the testee's knowledge on the test instruments 7 and 8.
5. Analyse deficiency of classic tests, two test items were developed with the dimensions of knowledge in evaluating and cognitive processes that contain the procedural knowledge of the testees on the test instruments 9 and 10.

Furthermore, in the second basic competency, Analysed Item Response Theory keep 5 performance indicators are developed as follows.

1. Analysed the concepts of modern test theory, two test items were developed with the dimensions of knowledge in analysing and cognitive processes containing the conceptual results of testee knowledge on the test instruments 11 and 12.
2. Analysed the concepts of modern test theory, three test items were developed with the dimensions of knowledge in analysing and cognitive processes which contained the conceptual results of testee knowledge on the test instruments 13, 14 and 15.
3. Analysed the elements of a modern test, developed three test items with dimensions of knowledge in analysing and cognitive processes that contain procedural knowledge of testees on test instruments 16, 17 and 18.
4. Analysed the objectives of the grain response theory, two test items were developed with the dimensions of knowledge in analysing and cognitive processes that contain the conceptual results of testee knowledge on test instrument items 19 and 20.
5. Analysed the item response theory assumptions, developed two test items with the dimensions of knowledge in analysing and cognitive processes that contain the conceptual results of testees knowledge on test instrument items 21 and 22.

Then in the third basic competency, Analysed the HOTS concept keep 5 performance indicators are developed as follows.

1. Analysed the concept of drafting the test, two test items were developed with the dimensions of knowledge in analysing and cognitive processes containing the conceptual results of testee knowledge on test instrument items 23 and 24.
2. Arrange the clue question, two test items were developed with the dimensions of knowledge in analysing and cognitive processes containing metacognitive knowledge of the testees on the test instruments 25 and 26.
3. Arrange test items, developing two test items with the dimensions of knowledge in evaluating and cognitive processes that contain procedural knowledge of the testees on test instrument items 27 and 28.
4. Analysed the test items with the question clue, developed two test items with the dimensions of knowledge in analysing and cognitive processes that contain procedural knowledge of the testees on test instrument items 29 and 30.
5. Assessing instrument test, two test items were developed with the dimensions of knowledge in analysing and cognitive processes that contain procedural knowledge of the testees on the test instruments 31 and 32.

In the fourth basic competency, Analysed Models Item Response keep 5 performance indicators are developed as follows.

1. Analysed 1 logistic parameters (Difficulty), two test items were developed with the dimensions of knowledge in analysing and cognitive processes

containing the conceptual results of the knowledge of the testees on the instruments 33 and 34

2. Analysed 2 logistic parameters (Difficulty & Discrimination), four test items were developed with the dimensions of knowledge in analysing and cognitive processes containing the conceptual results of the testee's knowledge on the test instruments 35, 36, 38 and 43.
3. Analysed 3 logistic parameters (Difficulty, Discrimination & Guessing), developed three test items with the dimensions of knowledge in analysing and cognitive processes that contain procedural knowledge of the testees on test instrument items 37, 39 and 40.
4. Analysed 4 logistic parameters (Difficulty, Discrimination, Guessing & Carelessness, two test items were developed with the dimensions of knowledge in analysing and cognitive processes containing the conceptual results of the testee's knowledge on the test instrument items 41 and 42.
5. Counting IRT models, two test items were developed with the dimensions of knowledge in analysing and cognitive processes which contained the conceptual results of the knowledge of the testees on the items 44 and 45 of the test instruments.

In the fifth basic competency, Determine HOTs Instrument in IRT keep 2 performance indicators are developed as follows.

1. Determine HOTs Instrument, four test items were developed with the dimensions of knowledge in analysing and cognitive processes containing procedural knowledge of the testees on the test instrument items 46, 47, 48 and 49.
2. Counting HOTs Instrument in IRT, three test items were developed with the dimensions of knowledge in analysing and cognitive processes containing procedural knowledge of the testees on the test instruments 50, 51 and 52.

The following shows the table 2 of results of developing competency achievement indicators which is used as a reference for the preparation of test items.

TABLE II. THE HOTS QUESTION INDICATORS

Basic Competencies	Indicators HOTS Competency Achievement	Cognitive Process and Knowledge Dimensions	No. Items
3.1 Analyzed Classical test theory	3.1.1 Analyzed classical test concepts and methods	C5K2	1, 2
	3.1.2 Analyzed classic test requirements	C4K3	3, 5
	3.1.3 Analyzed the concept of calculating validity and reliability	C4K2	4, 6
	3.1.4 Analyze the concept of item difficulty index		
	3.1.5 Analyze deficiency of classic tests	C5K2	7, 8
		C5K3	9, 10
3.2 Analyzed Item Response Theory	3.2.1 Analyzed the concepts of modern test theory	C4K2	11, 12
	3.2.2 Analyzed the concepts of modern test theory	C4K2	13, 14, 15
	3.2.3 Analyzed the elements of a modern test	C4K3	16, 17, 18
	3.2.4 Analyzed the objectives of the grain response theory	C4K2	19, 20
	3.2.5 Analyzed the item response theory assumptions	C4K3	21, 22
3.3 Analyzed the HOTS concept	3.3.1 Analyzed the concept of drafting the test	C4K2	23, 24
	3.3.2 Arrange the clue question	C4K4	25, 26
	3.3.3 Arrange test items	C5K3	27, 28
	3.3.4 Analyzed the test items with the question clue	C4K3	29, 30
	3.3.5 Assessing instrument test	C4K3	31, 32
3.4 Analyzed Models Item Response	3.4.1 Analyzed 1 logistic parameters	C4K2	33, 34
	3.4.2 Analyzed 2 logistic parameters	C4K2	35, 36, 38, 43
	3.4.3 Analyzed 3 logistic parameters	C4K3	37, 39, 40
	3.4.4 Analyzed 4 logistic parameters	C4K2	41,42
	3.4.5 Counting IRT models	C4K2	44, 45
3.5 Determine HOTS Instrument in IRT	3.3.2 Determine HOTs Instrument	C4K3	46, 47, 48, 49
	3.5.2 Counting HOTs Instrument in IRT	C4K3	50, 51, 52

C. Validation and Evaluation Stage

At this stage, it begins with validating the product design by asking experts in the field of the content or test material. The research team consisting of lecturers / evaluation experts validates the product with a technical panel by providing the latest form and review, so that it will produce an evaluation of comments and suggestions regarding whether an item is essential or relevant in product development. Comments and suggestions from experts are used to improve and revise the product being developed.

From the results of the analysis of the validity level of the items with the criteria for higher order thinking skills and question construction, it can be concluded that there are 43 items (83%) that are valid both in terms of material aspects, HOTS question criteria, and question construction (questions no.1, 2, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 24, 25, 26, 27, 28, 31, 32, 34, 35, 36, 37, 38, 39, 40, 42, 44, 45, 46, 47, 48, 49, 50, 51, 52) and 9 items (17%) were declared discarded which consisted of: invalid material (questions no.29, 33, 43); invalid from HOTS criteria (item 3, 5, 23, 41);

and invalid from the material and criteria for HOTS questions (questions no. 21, 30).

Test items that pass the expert validation are then tested for empirical validation on limited trials and large group trials. In a limited trial, the test items that were declared valid in the expert validation test were tried out on 15 students of PPs. Undiksha. This test is intended to measure the quality of the test items on a limited scale which includes measuring the validity, reliability, level of difficulty, distinction, and effectiveness of deceivers as well as to obtain student responses to the test items which will be used as a reference for revising and continuing the test items to the next stage.

Based on the criteria for drawing conclusions from the overall results of the following limited trial analysis: (1) accepted, if the test items are declared valid with the material, valid with HOTS test criteria, and valid with multiple choice test construction, (2) revision, if the test items are stated valid with material, valid with HOTS test criteria, and invalid with multiple choice test construction, and (3) discarded, if the test items are declared invalid with the material, invalid by HOTS test criteria, and invalid with multiple choice test construction. It can be concluded that there were 31 test items (72%) that passed, namely test item no. 1, 3, 4, 5, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 23, 25, 27, 28, 29, 32, 33, 34, 36, 38, 39, 40, 41, 42, 43, there were 3 test items (7%) declared to have passed with the revision of the cheat, namely test item no. 9, 24, 31, and 9 test items (21%) were declared null and void, namely test item no. 2, 6, 16, 21, 22, 26, 30, 35, 37.

The implementation of large group trials is intended to know more deeply the quality of the test items from the results of the quantitative analysis of validity, reliability, difficulty level, distinction, and level of distraction by testing test items on a large enough sample. The determination of the number of samples is adjusted to Nunanly's opinion that the number of research samples is 5 to 10 times the number of test items to be developed or analysed (Koyan, 2014). In this stage of the large group trial, the number of test items to be developed or analysed was 34 test items after 9 test items were declared invalid on expert validation and 9 test items were declared invalid in the small group trial with a total sample size of 5-10 x 34. In this large group trial, the standard number used in determining the number of research samples was 5 x 34 = 170, so that the number of research samples in this large group trial was 30 students who were included in the population.

The criteria for drawing conclusions from the entire series of large group trial analysis are as follows: (1) passes, if the test item is declared valid, the difficulty level of the test item is moderate, the minimum difference is not good, and all the trickster is effective, (2) passes with revision , if the test items are declared valid, the difficulty level of the test items is moderate, the minimum difference is not good, and there are ineffective distractors, and (3) fails, if the test items are declared invalid, the difficulty level of the test items is easy and difficult, the difference is negative (very bad), and there are ineffective trickster. The conclusions can be formulated as follows. There were 19 test items (56%) that were believed to have passed, namely test item no. 5, 6, 8, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22, 23, 24, 26, 28, 30, 31, there are 3 test items (9%) passed with the cheat revision namely test item no. 9, 24, 31, and there were 12 test items (35%) declared invalid, namely test item no. 1, 2, 3, 7, 9, 10, 15, 25, 27, 29, 33, 34.

D. Final Product Design Stage

Based on the results of the analysis of expert validation tests, limited trials, and large group trials, the following conclusions can be drawn: (1) of the 34 test items that passed the limited trial and continued to the large group trial, 22 test items were stated. passed and can be used which consists of 19 test items declared to have passed the entire quantitative analysis (test items nos. 5, 6, 8, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22, 23, 24 , 26, 28, 30, 31), (2) 3 test items that did not pass only in the cheater effectiveness analysis, the ineffective fraud revision was carried out (test items no.9, 24, 31), and (3) there were 12 items. The test was declared invalid because it did not pass two or more quantitative analyses, namely test item no. 1, 2, 3, 7, 9, 10, 15, 25, 27, 29, 33, 34. So, the HOTS-based *classroom assessment* test that was developed consists of 22 items.

IV. CONCLUSION

Based on the results of the study, several conclusions can be drawn as this point.

- The steps for developing a test for HOTS-based *classroom* assessment in include PPs. Undiksha the preparation stage which consists of gathering initial information by analysing Basic Competencies (KD), the initial design stage which consists of developing competency attainment indicators in the grid and designing the initial design. the test, the validation and evaluation stage consisting of expert validation, limited trials, initial product revisions, large group trials, and final revisions, and the final product design stage, which consists of designing the final test design.
- The quality of the test of HOTS-based economic subjects for class XI SMA based on the results of the analysis qualitatively and quantitatively, it can be stated that the test kits developed have good quality with content validity based on the results of expert validation that of the 52 test items developed there are 43 (83%) the test items are declared valid and relevant to be tested. Meanwhile, empirical validity resulted in 41 (95%) of the 43 test items developed that were declared valid and based on the results of large group trials resulted in 24 (71%) of 34 developed test items declared valid. Furthermore, the quality of the test in terms of the level of reliability of the test shows that in limited trials it produces a reliability value of 0.928 (very high) and in large group trials it produces a reliability value of 0.640 (high). However, in terms of the difficulty level of the items, it shows that in the limited trial of the 43

test items developed there were 5 (12%) items in the easy category, 37 (86%) the test items were in the medium category, and 1 (2%) the test items were categorized as difficult. (test item no.30), as well as in large group trials of the 34 test items developed there were 2 (6%) easy test items, 21 (62%) medium test items, and 11 (32%) test items categorized difficult. In terms of the difference in the test items, it shows, in the limited trial of the 43 test items developed there were 1 (2%) test items with a negative difference value, 5 (12%) test items that had a poor difference, 19 (44%) test items with good enough difference, and 17 (40%) test items with good differentiation power, and in the large group trial of the 34 test items developed there were 4 (12%) test items with negative or very bad difference values, 18 (53%) test items with poor distinction, and 12 (35%) test items with good enough difference. Finally, in terms of the effectiveness of cheaters, it shows, in the limited trial of the 43 test items developed there were 33 (77%) test items with all of the trickster effective, 8 (19%) test items with 1 intrigue were not effective, and there were 2 (5%)) test items with 2 distractors are ineffective, and in the large group trial of 34 test items developed there were 26 (77%) test items with all of the trickster being effective, 3 (9%) test items with 1 distractor were not effective, and there were 5 (14%) test items with 2 cheaters were not effective.

REFERENCES

[1] Kemendikbud, Panduan Implementasi Kecakapan Abad 21 Kurikulum 2013 di Sekolah Menengah Atas. Jakarta: Direktorat PSMA, 2017.

[2] Y. Abidin, Revitalisasi Penilaian Pembelajaran dalam Konteks Pendidikan Multiliterasi Abad Ke-21. Bandung: PT Refika Aditama, 2016.

[3] L. Lewy, Z. Zulkardi and N. Aisyah, "Pengembangan soal untuk mengukur kemampuan berpikir tingkat tinggi pokok bahasan barisan dan deret bilangan di kelas IX akselerasi SMP Xaverius Maria Palembang," Jurnal Pendidikan Matematika, vol. 3, no. 2, pp. 14-28, 2009.

[4] Kemendikbud. Permendikbud Nomor 24 Tahun 2016 tentang Kompetensi Inti dan Kompetensi Dasar Pelajaran pada Kurikulum 2013 pada Pendidikan Dasar dan Pendidikan Menengah.

[5] H. Stiadi, "Pelaksanaan Penilaian pada Kurikulum 2013," Jurnal Penelitian dan Evaluasi Pendidikan, vol. 20, no. 2, pp. 166-170, 2016.

[6] M.D. Gall, W.R. Borg and J.P. Gall, Educational research introduction (6th ed.). White Plains, NY: Longman Publishers USA, 1996.