Journal of Education and Practice www.iiste.org
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online) DOI: 10.7176/JEP
Vol.10, No.15, 2019

The Development of Higher Order Thinking Skills-Based Assessment Instrument for Elementary School Integrated Thematic Learning

Komang Okayana* Irawan Suntoro Lilik Sabdaningtyas Darsono
Faculty of Teacher Training and Education, University of Lampung
Soemantri Brojonegoro St. No.1 Gedung Meneng Bandar Lampung Indonesia 35145

Abstract
This study was aimed at developing a higher order thinking skills (HOTS)-based test instrument that is empirically and theoretically feasible for elementary school integrated thematic learning. This study was conducted through a Research and Development methodology (Borg & Gall, 1983) to a population of an elementary school fourth-grade students in Central Lampung by using a purposive sampling. A total of 64 participants took part in this study. The data were collected through questionnaires and tests. The results show that the test instrument developed was theoretically feasible with an average expert score of 90.14. This fell into very good category and was empirically feasible. A total of 29 questions were valid and internally consistent with a moderate level of difficulty, good discrimination power, and good distractor.
Keywords: higher order thinking skills (HOTS), assessment, integrated learning, thematic learning
DOI: 10.7176/JEP/10-15-16
Publication date:May 31st 2019

1. Introduction

It is generally accepted that education plays a very significant role in this 21st century. To be more competitive, the 21st century skills are highly required by students for achieving goals to make a change (Partnership for 21Century Learning; National Science Foundation; Trilling & Fadel, 2009). On the contrary, the 21st education also requires students to be more literate and able to think more critically that they do not fall into the trap of a hoax (Nugroho, 2018). Therefore, various skills which include creativity, critical thinking, communication, and collaboration (4Cs) to face this century should be well prepared (Partnership for 21st Century Learning; National Science Foundation; Hanifah, 2019; Kamarudin, 2016). In the context of education, classrooms are the best place where teachers can transfer the skills. However, in reality there is a distant gap between the so-called 21st demands and the implementation of learning carried out by teachers in the classrooms. Teachers have not trained their students to think and work on test questions according to the 21st century skills. The skills can be trained to students by way of implementing the 4Cs-oriented test questions by using a higher order thinking skills (HOTS)-based assessment instrument (RAND Corporation, 2012).

There is now much evidence to support the importance of HOTS to achieve goals through the ability to think critically, creatively, systematically, collaboratively, and communicatively to face this century (Weiss. E, 2003; Mirii, et.al., 2007) and HOTS is proved to be able to help students improve the skills (Syarifah, 2018; Handayani, 2013). Through the exposure of HOTS-based questions, students can be trained to have various skills such as problem solving, critical thinking, creative thinking, reasoning, and decision making (Resnick, 1987; Fanani, 2018; Aboesalem 2016). However, little is known about the implementation of HOTS-based questions in elementary school level.

Based on the author's observation and needs analysis, 30 class teachers in Seputih Banyak district, Central Lampung, made their questions conventionally. In other words, the questions were not based on the needs of the 21st century. They did not develop a HOTS-based test instrument, either. They also said that they had not participated in an instrument development training, 84% of them had not provided their students with a material review prior to exams, and 100% of them had not analyzed the question items made. In addition, based on the author's items analysis of their mid-term and final examination questions, it was found that the questions were not able to stimulate the students to think critically, creatively, communicatively, and collaboratively. In other words, there is a lack of questions to stimulate and train students to have the ability of the 4Cs competences. Thus, if this situation continues to exist, the learning goals of the century are nearly impossible to achieve. Hence, an additional study to deal with this situation was needed.

2. Research Methodology

This study was conducted by carefully following steps proposed by Borg & Gall's (1983) research and development (R & D) methodology for developing quality and effective educational products. The process of following the steps are as follows.

2.1 Needs analysis and identification of problems
In this first step, several problems such as conventional question items made by teachers, teachers' lack of understanding of constructing HOTS-based assessment instruments, teachers' lack of items analyses were identified. However, some potential supports to conduct this study such as teachers' support and representative school facilities and infrastructures were also identified.

2.2 Data Collection and Product Planning
At this stage, an in-depth analysis of Curriculum 2013 which was the one applied at the school was carefully done followed by analyzing the basic competences, constructing a HOTS grid, choosing a stimulus, making questions, making an answer key, and constructing a guideline for scoring.

2.3 Preliminary Product Design
A preliminary product design or prototype was made at this stage. It was undertaken based on the concept of HOTS by referring to the grid that was made.

2.4 Product Design Validation
At this stage, the product prototype was then assessed and evaluated by experts using a questionnaire to see the feasibility of the instrument. The validators consisted of assessment experts, material development experts, and linguistic expert. The advice and suggestions from the experts were then used to revise the Prototype I. The feasibility analysis was obtained using the following formula.

$$\text{Final Score} = \frac{Score\ obtained}{Maximal\ score} \text{ x } 100$$

The final score was converted into the following category as shown in The table 1 below.

Table 1. Conversion of Experts' scores

Score Interval	Category
76 – 100	Very Good
51 – 75	Good
26 – 50	Sufficient
0 – 25	Poor

2.5 Product Revision
The product revision was performed based on the experts' suggestions. After the changes were made based on the experts' feedback, the Prototype II was then developed followed by the main field test.

2.6 Small Classroom Experiment
At this stage, the Prototype II was tried-out at public elementary school 3 Swastika Buana in Central Lampung with a total sample of 20 (twenty) students. It was to find out the validity, reliability, level of difficulty, discrimination power, and the effectiveness of distractors of each item. After that, the questions considered valid were used and those of invalid were dropped. After such revision was done, the Prototype III was then developed. Then, this Prototype III was tested in a larger classroom.

2.7 Large Class Experiment
At this stage, the Prototype III was tested to 44 students at the same elementary school. The results of the test were then analyzed to find out the validity, reliability, level of difficulty, discrimination power, and the effectiveness of the distractor of each item. After that, the Prototype IV was then finally developed. This prototype IV was the final product of the development of this test instrument.

3. Results and Discussion

Based on the results, it was found that the question items made by the teachers were not in accordance with the demands of the 21st century skills. The teachers did not make a materials review prior to exams. They neither made HOTS-based assessment instrument nor undertook the items analysis. However, an assessment is theoretically in need to be done by teachers for measuring how far students have comprehended the learning materials delivered by the teachers (Hosnan, 2014), in which the results of the assessment can be used to decide the students' competence or ability and their learning achievement (Kankam Boadu, et al., 2015).

The findings of this study are in line with findings found by Nova, et. al., 2016; Budiman & Jaelani, 2014 that an assessment instrument needs to be tested in order to obtain theoretical and empirical feasibility. The theoretical feasibility test was carried out by three experts including assessment experts, material development

experts, and linguistic experts. To find out the empirical feasibility, it was tested to students in which the results of the test were then analyzed to find out the validity, reliability, level of difficulty, discrimination power and effectiveness of distractors in the form of multiple choices. Novitasari N. et.al, (2015) also explains that an assessment instrument needs to be tested in order to obtain theoretical and empirical feasibility.

The development of the test instrument in this research refers to Borg & Gall (1983) with the following steps.

3.1 Needs analysis and identification of problems

Several problems such as conventional question items made by teachers, teachers' lack of understanding of constructing HOTS-based assessment instrument, teachers' lack of items analysis were identified. However, some potential supports to conduct this study such as teachers' support and representative school facilities and infrastructures were also identified.

3.2 Data Collection and Product Planning

An in-depth analysis of Curriculum 2013 which was the one applied at the school was carefully done followed by analyzing the core and basic competences.

Table 2. Core and Basic Competences

Core Competence	3. Understanding factual and conceptual knowledge by observing and asking questions based on curiosity, God's creatures and activities, and objects that are found at home, at school, and at the playground.
Basic Competence	3.4 Connecting forces with motion in environmental events (natural sciences) 3.5 Identifying economic activities and their relationship with a variety of professions as well as social and cultural lives in the surrounding to the province (social sciences) 3.6 Comprehending fictional characters (Indonesian language) 3.7 Having knowledge of local dance motions (Arts) 3.8 Explaining the benefits of various individual characteristics in daily life (Civic education)

After that, it was then followed by constructing a HOTS grid, choosing a stimulus, making questions, making an answer key, and constructing a guideline for scoring.

3.3 Preliminary Product Design

A preliminary product design or prototype was made at this stage. It was undertaken based on the concept of HOTS by referring to the grid that was made. Then, the Prototype I was developed.

3.4 Product Design Validation

At this stage, the product prototype was then assessed and evaluated by experts using a questionnaire to see the feasibility of the instrument. The validators consisted of assessment experts, material development experts, and linguists. The advice and suggestions from the experts were then used to revise the Prototype I and to state that the design of the test instrument was feasible.

Tabel 3. Experts' Validation

No.	Evaluation Expert's Advice	Revision Results
1.	Questions with "except" statements must be underlined or typed in bold.	As advised
2.	The use of the preposition "at" must be adjusted if it is used to refer to a place or with a verb.	As advised
3.	Questions must be adjusted to the HOTS indicators.	As advised
4.	The choice of answers must vary and not use repeated words	As advised
	Material Expert's Advice	
1.	The indicators of the formulated questions should be much richer than those of the basic competence.	As advised
2.	The questions should be adjusted to the HOTS characteristics.	As advised
3.	The questions should be adjusted to the students' or school's location.	As advised
4.	The distribution of questions should be adjusted to the related material or basic competence indicators.	As advised
5.	A measure of the question request is needed.	As advised
6.	A rational distribution of questions is also required.	As advised
	Linguistic Expert's Advice	
1.	The use of the preposition "at", "to" should be in accordance with the standardized Indonesian language.	As advised
2.	The writing of the answer choices should be adjusted. If it is in the beginning of a sentence, it is printed in capitals. If it is in the end, use a period (.).	As advised
3.	Proper names should be printed in capitals.	As advised
4.	The imperative sentences should be provided with a "!" symbol at the end of the sentences.	As advised

The results of the experts' validation fall into very good category as shown in the Table 4 below.

Table 4. Results of Experts' Validation

No.	Expert	Score 1	Score 2	Average	Category
1.	Evaluation expert	94.12	100	97.06	Very Good
2.	Material expert	66.67	100	83.34	Very Good
3.	Linguistic expert	80	100	90	Very Good

3.5 Product Revision

The product revision was performed based on the experts' suggestions. After the changes were made based on the experts' feedback, the Prototype II was then developed followed by the main field test in a small classroom. The results of the test at this stage are as follows.

3.6 Small Classroom Experiment

Multiple Choices

Table 5. Instrument Validity Test

Number of Questions	Total	Description
1, 2, 4, 5, 8, 9, 11, 13, 15, 16, 18, 20, 23, 25, 29, 32, 34, 36, 38, 40, 42, 45, 48, 49, 51, 52, 54, 57, 59, 60	30	Valid ($r_{value} > r_{table}$)
3, 6, 7, 10, 12, 14, 17, 19, 21, 22,24, 26, 27, 28, 30, 31, 33, 35, 37, 39, 41, 43,44, 46, 47, 50, 53, 55, 56, 58	30	Invalid ($r_{value} \leq r_{table}$)

To find out the validity of this instrument, a Product Moment Correlation analysis was run. An item is said to be valid if the r_{value} is higher than r_{table}. The score of the r_{table} in this group of questions is 0.444. Thirty (50%) questions are valid, while the other thirty are invalid.

Table 6. Instrument Reliability Test

Questions	r_{value}	Criteria
1-60	0.954	Very High

Table 7. Questions Level of Difficulty

Category	Number of questions	Total
0.71 – 1.00 (Easy)	3, 6, 7, 10, 12, 14, 17, 19, 21, 22, 24, 26, 27, 28, 31, 33, 35, 37, 39, 41, 43, 44, 46, 47, 50, 55, 56	27
0.31 – 0.71 (Medium)	1, 2, 4, 5, 8, 9, 11, 13, 15, 16, 18, 20, 23, 25, 29, 32, 34, 36, 38, 40, 42, 45, 48, 49, 51, 52, 54, 57, 59, 60	30
0.00 – 0.30 (Difficult)	30, 53, 58	3

Table 8. Discrimination Power

Range	Questions	Category
0.40 – 1.00	1, 2, 4, 5, 8, 9, 11, 13, 14, 15, 16, 18, 20, 23, 25, 29, 32, 34, 36, 38, 40, 42, 45, 48, 49, 51, 52, 54, 57, 59, 60	Very Good
0.30 – 0.39	30	Good
0.20 – 0.29	17, 19, 24, 27, 43, 44, 55, 56	Sufficient
0.00 – 0.19	3, 6, 7, 10, 12, 21, 22, 26, 28, 31, 33, 35, 37, 39, 41, 46, 47, 50, 53, 58	Poor

Table 9. Question Distractors

Category	Questions	Total
rpbis positive answer key, Response >5%, and rpbis negative distractor	1, 2, 4, 5, 8, 9, 11, 13, 15, 16, 18, 20, 23, 25, 29, 32, 34, 36, 38, 40, 42, 45, 48, 49, 51, 52, 54, 57, 59, 60.	30
rpbis negative answer key, Response <5%, and rpbis positive distractor	3 ,6, 7, 10, 12, 14, 17, 19, 21, 22, 24, 26, 27, 28, 30, 31, 33, 35, 37, 39, 41, 43, 44, 46, 47, 50, 53, 55, 56,, 58.	30

Description

Table 10. Question Validity

Questions	Total	Description
2,4,5,10,11	5	Valid ($r_{value} > r_{table}$)
1,3,6,7,8,9,12	7	Invalid ($r_{value} < r_{table}$)

Table 11. Question Reliability

Questions	r_{value}	Criteria
1 – 12	0.787	High

Table 12. Question Level of Difficulty

Category	Questions	Total
0.71 – 1.00 (Easy)	1,3,7,9,12	5
0.31 – 0.71 (Medium)	2,4,5,10,11	5
0.00 – 0.30 (Difficult)	6,8	2

Table 13. Discrimination Power

Range	Questions	Category
0.40 – 1.00	2,4,5,10,11	Very Good
0.30 – 0.39		Good
0.20 – 029	1,6	Sufficient
0.00 – 0.19	3,7,8,9,12	Poor

Each item in the Prototype II was analyzed. Thirty multiple choice questions and five essay questions were considered empirically feasible because they were proved to be valid and reliable with a moderate level of difficulty and good discrimination power. The effectiveness of the distractors was also proved to be good. After the Prototype II was revised, then the Prototype III was developed which was then tested in a larger-size class.

3.7 The Result of Large Class Instrument Try-out

The Prototype III was tested to fourty four subjects. The results of the test were then analyzed to find out the validity, reliability, level of difficulty, discrimination power, and the effectiveness of the distractor of each item. The following are the results of the analyses.

Multiple Choices

Table 14. Question Validity

Questions	Total	Description
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30	30	Valid ($r_{value} > r_{table}$)
-	-	Invalid ($r_{value} \leq r_{table}$)

Table 15. Question Reliability

Questions	r_{value}	Criteria
1 – 30	0,892	High

Table 16. Level of Difficulty

Category	Questions	Total
0.71 – 1.00 (Easy)	8, 10, 11, 13, 14, 15, 16, 21, 25, 28	10
0.31 – 0.71 (Medium)	1, 2, 3, 4, 5,6, 7, 9, 12, 17, 18, 19, 20, 22, 23, 24, 26, 27, 29	19
0.00 – 0.30 (Difficult)	30	1

Table 17. Discrimination Power

Range	Questions	Category
0.40 – 1.00	2, 3, 5, 7, 8, 9, 12, 20, 21, 23, 26, 27, 28, 30	Very Good
0.30 – 0.39	1, 2, 6, 10, 14, 16, 17, 18, 19	Good
0.20 – 0.29	11, 15, 22, 24, 25, 29	Sufficient
0.00 – 0.19	13	Poor

Table 18. Distractor

Category	Questions	Total
rpbis positive answer key, Response >5%, and rpbis negative distractor	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30	30
rpbis negative answer key, Response < 5%, and rpbis positive distractor	-	0

Description

Table 19. Questions Validity

Questions	Total	Description
1,2,3,4,5,	5	Valid $r_{value} > r_{table}$
-	-	Invalid $r_{value} \leq r_{table}$

Table 20. Question Reliability

Questions	r_{value}	Criteria
1 – 5	0.635	Medium

Table 21. Question Level of Difficulty

Category	Questions	Total
0.71 – 1.00 (Easy)		-
0.31 – 0.71 (Medium)	1,2,3,4,5	5
0.00 – 0.30 (Difficult)		-

Table 22. Discrimination Power

Range	Questions	Category
0.40 – 1.00	1,2,4	Very Good
0.30 – 0.39	3,5	Good
0.20 – 0.29	-	Sufficient
0.00 – 0.19	-	Poor

Each item in the Prototype III was analyzed. Twenty-nine multiple choice questions and five essay questions were considered empirically feasible because they were proved to be valid and reliable with moderate level of difficulty and good discrimination power. The effectiveness of the distractors was also proved to be good. After the Prototype III was revised, then the Prototype IV as the final product was developed.

4. Conclusion

Based on the results and discussion, it can be concluded that the final product in this study is a HOTS-based assessment instrument that is theoretically and empirically feasible for the integrated thematic learning of elementary school fourth-grade students. The feasibility of the instrument was obtained from the experts' evaluation and instrument try-outs in the classrooms. This instrument has been theoretically feasible because it was validated by the assessment, material, and linguistic experts, in which the results fell into very good category. This multiple-choice instrument as well as the essay questions are empirically feasible because they were tested in classrooms. The results of the test were proved to be valid and highly reliable with a moderate level of difficulty and good discrimination power. The effectiveness of the multiple-choice distractors was also proved to be good.

Refferences

Abosalem, Yousef. (2016). Assessment techniques and students higher-order thinking skills. *International Journal of Secondary Education*, *4*(1): 1-11.

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

Budiman, A. Jailani. (2014). Pengembangan instrumen asesmen *higher order thinking skill* (HOTS) pada mata pelajaran matematika smp kelas VIII semester 1. Jurnal Riset Pendidikan Matematika *1*(2), 139-151.

Fanani. Moh. Zainal. (2018). *Strategi pengembangan soal higher order thinking skill (HOTS). Dalam: Kurikulum 2013.* Journal of Islamic Religious Education, *2*(1): 57-76.

Handayani. R & Priatmoko S. (2013). Pengaruh pembelajaran *problem solving* berorientasi HOTS (*higher order thinking skills*)terhadap hasil belajar Kimia siswa kelas X. Jurnal Inovasi Pendidikan Kimia. *7*(1). 1051-1062

Hanifah, Nurdinah. (2019). Pengembangan instrumen penilaian *higher order thinking skill* (HOTS) di sekolah dasar. *Current Research in Education: Conference Series Journal.* Vol. 1. No.1: 005.

Hosnan. (2014). Pendekatan saintifik dan kontekstual dalam pembelajaran abad 21: Kunci sukses kurikulum 2013. Jakarta: Ghalia Indonesia.

Kankam, B. et al. (2015). Teachers' perception of authentic assessment techniques practice in social studies lessons in senior high schools in Ghana. International Journal of Educational Research and Information Science. 10 Januari 2015. *1*(4), 62-68.

Miri, B., David, B. C., & Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. Research in Science Education, *37*(4), 353–369. doi:10.1007/s11165-006-9029-2

Nova, A.R, et.al. (2016). Pengembangan instrumen asesmen penguasaan konsep tes testlet pada materi suhu dan kalor. Jurnal Pendidikan: Teori, Penelitian, dan Pengembangan. *1*(6), 1197—1203.

Novitasari, et.al. (2015). Measuring problem solving skills of high school students on biology. Jurnal Biologi Edukasi. *7*(1), 1-6.

Nugroho, Arifin. R. (2018). *Higher order thinking skills*. Jakarta: Gramedia Widyasarana.

P21. (2013). *Reimagining citizenship for the 21st century: A call to action for policymakers and educators.* Washington DC, Partnership for 21st Century Skills.

RAND Corporation. (2012). Teaching and learning 21st century skills: Lesson from the learning sciences. Hong

Kong: Asia Society Global Cities Education Network.

Syarifah T.J. et.al. (2018). Higher order thingking (HOT) problems to develop critical thinking ability and student self efficacy in learning mathematics primary schools. National Seminar on Elementary Education. 1(1), 917-925.

The Partnership for 21st Century Skills. (2013). P21 common core toolkit, A guide to aligning the common core state standards with the framework for 21th century skills. Washington DC: The Partnership for 21st Century Skills.

Trilling, B., & Fadel, C. (2009). 21st century skills: Learning for life in our times. San Francisco: Wiley.

Weiss, E. (2003). Problem-based learning in the information age: Designing problems to promote higher order thinking. Wiley Periodicals, 95: 25-31.