Evaluating the Use of HOTS in the Formulation of English Teacher-Made Test

Tiara Novalia*, Suwandi**

* English Education, Post Graduate Program
** Universitas Negeri Semarang
Email: tnovalia11@gmail.com

DOI: 10.29322/IJSRP.12.09.2022.p12935
http://dx.doi.org/10.29322/IJSRP.12.09.2022.p12935

Paper Received Date: 3th August 2022
Paper Acceptance Date: 5th September 2022
Paper Publication Date: 15th September 2022

Abstract- Since implementing the 2013 curriculum, the government has focused on applying HOTS at school. Beginning in 2018, HOTS questions have been included in the national examination. However, in the national examination of 2018, 40 % of the students in Indonesia struggled to answer the HOTS questions. As a result, there is a decrease of 0,93 in the students' scores in both in-state and private high schools. This study focussed on evaluating the use of HOTS in English teacher-made tests at MA NU 03 Sunan Katong Kaliwungu Kendal. From the findings, the comparison between LOTS and HOTS was 47% to 53%. The proportion of HOTS was more than it is expected as the criteria of HOTS were expected to be around 10% to 15% of the total questions. While for the level distribution, analysing was 20%, evaluating was 6%, and the highest, creating with 74%. It suggested that the teacher had already given plenty of HOTS practice to her students in the class by using the daily test. The test also fulfilled the requirement of validity, practicality, and reliability, as shown in the supporting document.

Index Terms- Cognitive domain, Bloom's taxonomy, HOTS, English teacher-made test.

I. INTRODUCTION

The implementation of the 2013 curriculum creates the change in the approach, method, and learning assessment in Indonesia education system. The government tries to encourage the teachers, especially teachers of higher education level, to apply higher-order thinking skills in every subject at school. As a result, the government has involved HOTS questions in the national examination; as Muhajir Effendi (2018) states, beginning in 2018, HOTS questions have been included in the national examination. However, they are only 15 % out of the total question. Likewise, Collins (2014) states that applying HOTS can prepare students to compete in the 21st century. However, some issues are with HOTS in Indonesia (Wuryadi as cited in Rezkikasari, 2018).

According to Wuryadi (2018), Indonesia is left behind in applying HOTS to the students compared to other countries. It is supposed to be used years before. In addition, applying HOTS skills in the 2018 national examination does not go along with the reality of the teaching and learning process in the classroom. As a result, there is a disconnection between the national examination test makers and the students' conditions in the school (Wuryadi, 2018). The test makers do not know whether the teaching and learning process has been applied in HOTS or not. Still, some questions in the national examination are in HOTS form. As a result, most students have difficulties in doing the national examination. Based on the data from Kemendikbud, as stated by Totok as cited in Nugroho (2018), in the national examination of 2018, 40 % of the students in Indonesia struggled to answer the HOTS questions. As a consequence, the student's scores are dropped. Kemendikbud as cited in Nugroho (2018) states that there is a decrease of 0,93 in the students' scores, both in-state high school and private high school. Two significant factors cause it. First, the swift from paper-based to computer-based tests pushes the students to adapt to the test mode. Second, some questions in the form of HOTS make students struggle to answer the questions. Directorate of High School Development in the International Standard Preparation Guide explained that most high school teachers only tended to measure low-order thinking skills (LOTS). Teachers' questions commonly measured recall skills. In addition, teachers focused on theories, not contextual knowledge, which did not fit the 2013 curriculum's requirements.

Teachers are expected to promote HOTS elements to encourage more profound thinking activities in students. HOTS cannot be directly given to the student. However, students should develop their thinking skills if they get the assessment regularly. As a guide in the classroom, the teacher plays a big part in training HOTS to the students. Giving a good assessment is one of the best ways to train HOTS to the student. HOTS should be present in all the curriculum 2013, including the English language assessment. The assessment is not only used to recall, restate, and recite the information, but it should be more to make students analyse, synthesize, evaluate and create. The concept of the HOTS question refers to the abilities to (1) transfer one concept to another, (2) to process and apply the information, (3) to find the relevance of different information, (4) to solve the problem using information, and (5) critical thinking.

HOTS involves the transformation of information and ideas. This transformation occurs when students analyse, combine facts and opinions and synthesize, generalize, explain, or arrive at some conclusion or interpretation. Manipulating information and ideas through these processes allows students to solve problems, understand, and discover new meanings (Tomei, 2005). One of the most well-known taxonomies in education is Bloom's. It offers a basic thinking skills model adopted by several researchers for their studies' purposes. Bloom's taxonomy focuses on six levels of thinking that students practice while learning or acquiring knowledge. They are consisted of understanding, remembering, applying, analysing, evaluating, and creating. The first three levels of taxonomy belong to lower order thinking skills. While analysing, evaluating, and creating belong to higher order thinking skills. The focus point in this study is dealing with HOTS in Revised Bloom's Taxonomy. As the addition, this study examined the validity, practicality, and reliability of the test that used as the objects of the study according to Brown's (2004) theory on language assessment.

Validity is the most complex criterion and important principle of an effective test. There is no final absolute measure of validity, but several different kinds of evidence may be invoked in support (Brown, 2004). There are five types of evidence below, according to Brown (2004); content validity, criterion-related evidence, construct-related evidence, consequential validity, face validity.

Practicality refers to an effective test. This means that the test is not excessively expensive. It also stays within appropriate time constraints. Moreover, it is relatively easy to administer. Last, it has a scoring/evaluation procedure that is specific and time-efficient. Thus, it can be said that without those aspects, the test might be said as impractical. Practicality is determined by the teacher's and the student's time constraints, costs, and administrative details, and to some extent, by what occurs before and after the test, Brown (2004). It means that to administer what is so-called a good test, the teacher must meet some criteria relating to the practicality of the test.

Reliability means the consistency of measurement in the test. According to Brown (2004), "A reliable test is consistent and dependable. It means that if the teacher gives the test to a student on two different occasions, the test result will still be the same. The reliability problems may occur by the following factors as follows (Brown, 2004); student-related reliability, ratter reliability, test administration reliability, test reliability.

II. METHOD

The research was categorized as a qualitative study. Creswell (1998) stated that qualitative research involved the study in which it used and collected a variety of empirical methods. It analyzed the data based on particular theories in such a way to prove the evidence whether or not the data are compatible with those theories. Thus, the result was explained in the description.

This research aimed to explain the realization of HOTS in English teacher-made tests, so the content analysis was used as the research design. Hsieh and Shannon (2005) define content analysis as "a research method for the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns." Therefore, content analysis was chosen for some reasons; first, this study provided the readers with an explanation of the HOTS found in the test. Then, the examples of each level of HOTS were compared with the theories of revised Bloom's taxonomy. Last, the result of this study was written in description.

As the subject of the study, one of the senior high school teachers in an Islamic private school is chosen. She took part as an interviewee to provide information about her English test's practicality, reliability, and validity. Furthermore, the objects of this research were teacher-made test items for English subjects for two semesters from two academic years. Teacher-made tests are the tests that are constructed and prepared by the teacher themselves to test their students. Teachers made these kinds of tests by considering the students' abilities in the classroom. They also adjusted the material of the tests based on what students have learned in the classroom. In short, these tests might suit students because they knew what they did in the test and expected a better outcome. The objects were the school's documents of test paper that had been used for students in tenth, eleventh, and twelfth grade at MA NU 03 Sunan Katong Kaliwungu Kendal.

The reading and writing questions were selected as items to be analysed. There were two units of analysis used in this research. The first was a cognitive process in Revised Bloom's Taxonomy (Anderson et al., 2001) which contained six main categories and 19 sub-categories. The characteristics of each sub-category were used to examine the individual test item in which category the item was included. The second was the practicality, reliability, and validity of the test so that the test was appropriate to be given to the student.

The data of this research belonged to qualitative data. Qualitative data are mostly non-numerical and usually descriptive or nominal (Creswell, 2009). It means the data collected were in the form of words and sentences. In this study, the data were collected from documentation and interview. The documentation included teacher-made test items for the English subject for two semesters from two academic years of MA NU 03 Sunan Katong Kaliwungu Kendal.

In addition, interview is used as another method of data collection. We used the standardized open-ended interview to question the teachers. Patton (1990) stated that in a standardized open-ended interview, all interviewees were asked basic questions in the same order. Questions were worded in a completely open-ended format. We asked the teachers about the test's practicality, reliability, and validity here. The interview was personal, so we asked the questions directly to the interviewee. Then, the result was recorded and transcribed as the research finding. There were two kinds of instruments in this research. They were checklist forms and interview questions. Then, the checklist form was divided into four sub-forms. After preparing the instruments, we included teacher-made English subject tests. Then, we analysed the realization of HOTS, which is embedded in the test items. This large number of items was purposively selected to obtain a more accurate analysis. The steps of data collection are described below:

1) Selecting the test paper written by English teachers.
2) Grouping the test items based on the cognitive level categories; analysing, evaluating, and creating.
3) Observe the supporting documents for the test's practicality, reliability, and validity.
4) Interviewing the interviewee about the test evaluation
5) Transcribing the result of the interview reciprocal

The analysis aimed to investigate the quality of the items that agreed with the thinking skills category. We used the approach of Singh and Shaari (2019) to add the modification as they had done quite similar research. The analysis of items was categorized into three parts. First, we evaluated the items according to HOTS listed in Bloom's Taxonomy. These items were reading and writing tests. Second, we analysed the levels of HOTS in Bloom's Taxonomy. The items were then categorized into three levels; analysing, evaluating, and creating. In the next step, we categorized the sub-skills under each main skill; analysing, evaluating, and creating items. Then, we observed the supporting documents to check the test's practicality, reliability, and validity. Last, we identified the evaluation of the test based on the interview done with the teachers.

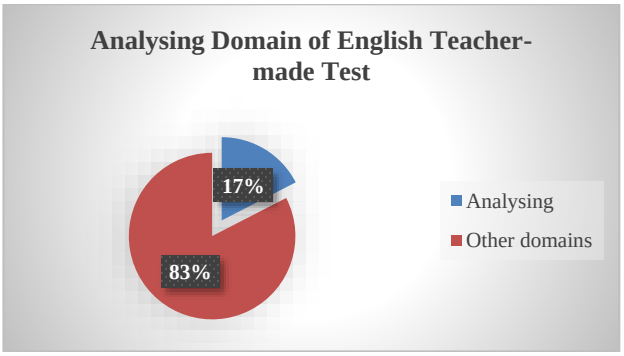
III. Results and Discussions

A. Analysing Domain in the Test

Based on Revised Bloom's taxonomy by Anderson (2001) that we used as a tool for analysing the test, the result was there were 241 questions belonged to LOTS and 212 to HOTS. The overall questions that we found in the question was 453 questions. The gap was only 6 %. It suggested that the teacher give plenty of HOTS exercises to the students. It is in line with the teacher's second interview (interview#2, 11062022). The teacher stated that the result was as expected because they aimed to give HOTS questions to the students. It is not surprising since the teacher has already planned the test before.

In addition, we analysed the English teacher-made test to examine the use of analysing domain in the questions. Based on the documents obtained, we found that there were 79 questions out of 453 questions belonging to the analysing domain of Bloom's revised taxonomy. Here is the result.

Figure 1 Analysing Domain of English Teacher-made Test



From the figure above, 17 % of the questions belonged to the analysing domain. The number was pretty high since it was more than 10 %. The data showed that the teacher had provided students with enough exercises to deal with the analysing domain questions. Furthermore, we also discovered that the teacher used some types of active verbs in the questions.

Table 1 Operational Verbs Used in the Test under Analysing

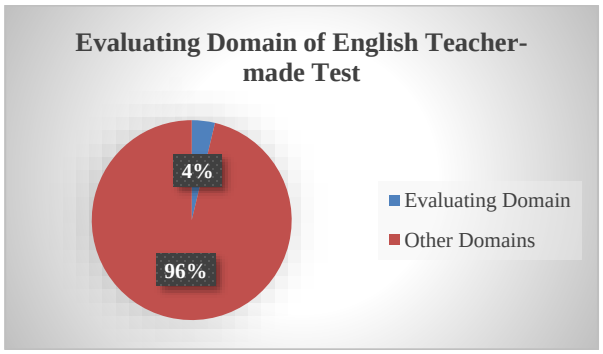
No	Operational Verbs	Number of questions
1.	Classify	56
2.	Analyse	10
3.	Identify	2
4.	Inference	5
5.	Predict	1
6.	Point out	4
7.	Compare	1

From the table above, it showed that the word "classify" appeared the most. The reason was in majority of the questions, the students needed to classify some kinds of text. For the second highest, analyse, appeared as the questions in which the students needed to analyse whether the options were true or false based on the text that they read.

B. Evaluating Domain in the Test

Based on the data, we found that there were 17 questions out of 453 questions belonging to evaluating domain.

Figure 2 Evaluating Domain of English Teacher-made Test



From the figure, evaluating domain only appeared in as much as 4 % of the questions. It is because evaluating domain was found mostly in reading questions. In contrast, none of them was found in writing questions. Moreover, the reading questions that used evaluating domain were in the form of essays. Meanwhile, there were only three questions in the form of a multiple-choice question. Furthermore, as written in the table, some active verbs under the evaluating domain were also found.

Table 2 Operational Verbs Used in the Test under Evaluating

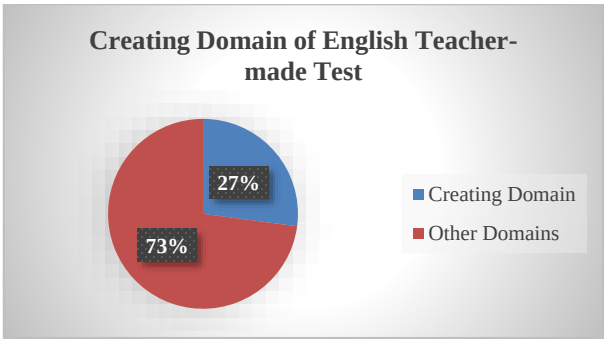
No	Operational Verbs	Number of questions
1.	Conclude	9
2.	Predict	1
3.	Evaluate	3
4.	Opinion	1
5.	Interpret	2
6.	Assess	1

Based on the table above, conclude appeared the most. Here, the word conclude was found in reading questions. The questions were about concluding the message of the text or song. To conclude something, the students needed to understand the text.

C. Creating Domain in the Test

The last type of higher-order thinking skill in the cognitive domain was creating. Of 453 questions, we found 116 questions that belonged to creating domain.

Figure 3 Creating Domain of English Teacher-made Test



The number was pretty high considering the other domains, such as analysing with 79 questions or evaluating with only 17 questions. The reason might be that the tests consisted of reading and writing questions, so almost half of the test was in writing. Most of the questions that belong to creating domain were in the form of essays.

Table 3 Operational Verbs Used in the Test under Creating

No	Operational Verbs	Number of questions
1.	Create	100
2.	Combine	10
3.	Arrange	6

HOTS questions that found majority in writing questions were creating. Perhaps, it was because the teacher wanted the students to improve more in creating or making something. However, so far, the implementation of HOTS in each test is beyond than what this study expected so far. The teacher gives students with lots of practice. Although it might be quite difficult at first, but the students can do more practice. It will benefit them in doing such a test in the future.

In the end, these findings are similar to the result of the research conducted by Narwianta, Anggani, and Rukmini (2019). The study entitled "The Evaluation of Higher Order Thinking Skills in English School National Standardized Examination at State Senior High School 6 Semarang". The results indicated that HOTS realized in the listening, reading, and writing questions. The listening was in the form of spoken and written. There is one listening question categorized into HOTS of level analysing. In reading, the questions belonging to HOTS reach eight questions that consist of 5 questions of analysing level and three questions of evaluating level.

The result shows that in reading questions, HOTS domains that appear in the test mostly analyse and evaluate. It is in line with the result that we found in the test. I found that the majority of reading questions belonged to analysing. From the three domains of HOTS, analysing tends to be the highest HOTS that appeared in the test. In addition, writing questions that appeared in the test is in creating domain. The finding of the research from Narwianta, Anggani, and Rukmini (2019) also suggests the same thing. They found that one HOTS question was the level of creating in writing questions. It is similar to the findings that I found. We discovered that writing questions were in the form of creating. While comparing HOTS in reading and writing questions, it seems that reading appears slightly more than writing questions.

In accordance to the validity, practicality, and reliability of the test. We followed Brown's (2004) theory on language assessment. We did some checklist and checked the supporting documents to see whether they followed the rules or not. From the data in findings, it can be said that the test is valid enough. The teacher has already met the criteria suggested by Brown (2002). The teacher has stated the objective of the test clearly in the test, has set the timing appropriately, and has created the test's structure well. The tests also are established in reference to a specific purpose; the test may not be valid for different purposes. For example, the test the teacher created used to evaluate the students' knowledge about pronouns may not be valid for predicting the students' listening skills. Nevertheless, it is one of the principles of assessment. It also leads to the following principle of assessment.

Similarly, a test's validity is established in reference to specific groups. These groups are called reference groups. The test may not be valid for different groups. For example, a test designed to measure the English of twelfth-grade students will not be appropriate to be given to tenth-grade students since the test materials are not the same. The teacher is responsible for describing the reference groups used to develop the test. Furthermore, the validity of the test in this research is measured by using the content validity method. Content validity refers to the extent to which the items on a test are reasonably representative of the entire domain the test seeks to measure. Content validity in this research is assessed by using Brown's theory.

We discover the practicality of the test based on Brown's (2004) theory of practicality. From the aspects of practicality, we can say that the test was practical enough. The teacher met most of the

requirements. I also delivered the result of the practicality test to the teacher. I said that there were two aspects that seemed not to be stated in the checklist. After conducting the interview, the teacher said that she thought all the students' activities needed some costs, especially when they have a test. Because the test itself is planned in school budgeting, for example, for copying the test sheets and so on, while for the method for reporting, there was a method determined in advance, for example, when it was the daily test, then the score would be combined to other scores to yield the final score which was written in students' report card.

Moreover, the test itself is well-structured. The teacher has provided the objective, the guides, the time allocation, and so on. The test is also comprised of five to ten questions so that it will be easier to score. The teacher also provides the scoring rubric. It is in line with the statement of Manuel (2022). He says that no matter how valid or reliable a test is, it has to be practical to make and to take this means that the test is economical to deliver. It means that it is not excessively expensive. The layout should be easy to follow and understand. The students should be able to follow the directions of the test. In addition, it should stay within appropriate time constraints. It means that it has to follow the time allocation. It should not waste much time than the allocation. Last, it is relatively easy to administer. Its correct evaluation procedure is specific and time-efficient.

Overall, the test is easy to administer. The teacher has set the time for each test to be given. The test also costs less. Since the test is only one page, the cost of a copy will be inexpensive. In addition, the teacher has already provided the scoring or evaluation procedure to make it easier to score the test. Indeed, we can say that the test is practical enough.

Moving to reliability, it refers to the consistency of the test. It means that if the test is given to the student in any other condition now or then, the result of the test will not change. In this research, reliability is limited to the theory of Brown (2004) on the reliability of the test, not quantitative.

The test has met 3 of 5 criteria based on Brown's checklist. The test does not meet all the requirements. However, it is still reliable because of the supporting document's evidence. The supporting documents are also available to support the data. For example, the scoring rubric is the supporting documentation to support the statement 'Is the objective of scoring procedure clear?'. It means that the test is easy to score since the scoring sheet is provided. The students also got each text copy as stated in the checklist.

While we did the data analysis, we found some tests the teacher gave the students in two different academic years. It means that the same test is given to different students repeatedly. Therefore, it can be concluded that the test managed to yield the same result, so the teacher gave the test again to the other students.

IV. Conclusion

There are some major conclusions that can be drawn from this study. First, the questions in the tests are mostly reading and writing, with the total questions almost similar in number. It might be better for the teacher to give students daily tests in other English skills, such as grammar or listening. On the other sides, the variation of reading and writing questions is good. The teacher provides the questions in the form of essays, fill-in, and multiple choices, so the students can practice doing such forms of questions. Second, the realization of HOTS in the tests is various. The teacher provides lots of HOTS questions. The number of HOTS is also similar to the number of LOTS found in the tests. It means that the teacher is doing a great job in providing the students with HOTS practices. It may help the students to solve HOTS questions in the future, for example, when they do the final examination or school examination. The students will be more prepared because they have done the same thing in their daily tests. Third, the operational verbs and the level of HOTS found in the tests are varied. The teacher uses the three levels of HOTS domains. Surprisingly, the creating level is highly used in the tests. Most of the questions are asked the students to create or to make sentences or text. It is good for the students because they are allowed to write something. It means that the teacher is aware of improving the students' writing skills. The teacher uses different kinds of operational verbs for each level. It can help the students to enrich their knowledge in dealing with HOTS questions. Unfortunately, for creating domain, the teacher mostly used the word creating as the operational verb. At the same time, other operational verbs can be used. If the others are used, the students will be exposed to more variations of the questions.

While for the test's validity, practicality, and reliability, they align with Brown's (2004) theory. The tests meet almost all the requirements of those three aspects. Although some aspects are not suited, the teacher is still trying to achieve the other aspects in terms of validity, practicality, and reliability of the tests.
Overall, the teacher's use of HOTS in the tests is done well. The teacher provides more than enough HOTS questions to students.

We also proposed some suggestions for the teacher, the students, and future researchers. First, what the teacher does in the tests is great. However, it will be more excellent if the teacher provides the tests in other English skills. In addition, the teacher can also give the students with more variety of questions. It will be great for the students to practice such a test. Second, the students can try to do the test as best as they can. It is their time to practice HOTS as their daily test before they do the HOTS test in their final examination. Third, future researchers can do a deeper analysis of a similar topic. The lack of this research is not using quantitative measurement to analyse the validity, practicality, and reliability. It will be better if a similar topic of this research is done with a mix of qualitative and quantitative methods.
.

References

[1] Anderson. L. W., Krathwohl. D. R, A taxonomy for learning, teaching, and assessing. London: Longman, 2001.

[2] Brown. H. D, Language assessment: principles and classroom practices. London: Longman, 2004.

[3] Creswell. J. W, Qualitative inquiry and research design, New York: Sage Publications, 1998.

[4] Hsieh. H. F., Shannon. S. E, “Three approaches to qualitative content analysis," *Qualitative Health Research*, 2005.

[5] Manuel. J, Testing: reliability, validity, and practicality, 2022. Https://Englishpost.Org/Reliability-Validity-and-Practicality/.

[6] Narwianta. N., Anggani. D. B. L., Rukmini, D, The evaluation of higher order thinking skills in English school nationally standardized examination at State Senior High School 6 Semarang, *English Education Journal*, 2019.

[7] Nugroho. B. P, Mengenal hots penyebab soal unbk dikeluhkan begitu sulit, 2018. Https://News.Detik.Com/Berita/d-3975448/Mengenal-Hots-Penyebab-Soal-Unbk-Dikeluhkan-Begitu-Sulit .

[8] Patton. M. Q, Qualitative research and evaluation methods. New York: Sage Publications.

[9] Singh. R. K., Shaari. A. H, The analysis of Higher-Order Thinking skills in English reading comprehension tests in Malaysia. *Malaysian Journal of Society and Space*, 2019.

[10] Rezkikasari, I, Indonesia dianggap terlambat terapkan HOTS, 2018. Https://Republika.Co.Id/Berita/Pendidikan/Eduaction/18/04/18/P7cy6m328-Indonesia-Dianggap-Terlambat-Terapkan-Hots.

AUTHORS

First Author – Tiara Novalia, Post Graduate Student Universitas Negeri Semarang, tnovalia11@gmail.com.
Second Author – Suwandi, Universitas Negeri Semarang.

Correspondence Author – Author name, email address, alternate email address (if any), contact number.