AN ANALYSIS OF VALIDITY AND RELIABILITY OF COGNITIVE TAXONOMY ITEMS ON ENGLISH SUMMATIVE TEST AT SMA NURUL FALAH PEKANBARU

Murni Wahyuni Sianturi, Dr. Erni, S.Pd., M. Hum, Dr. Dahnilsyah, SS., M. A.
Email : murni.wahyuni4690@student.unri.ac.id , erni@lecturer.unri.ac.id , danil_71@yahoo.com
Phone number: 081391661946

Student of English Study Program
Language and Arts Department
Faculty of Teachers Training and Education
Riau University

Abstract: *This study aimed to analyze the quality of summative test that was conducted by the teacher at the eleventh grade of SMA Nurul Pekanbaru and was constructed as descriptive quantitative research. The sample consisted of 53 students in the second semester of the academic year 2020/2021 and was selected by using total sampling with the lottery method. Furthermore, the instrument of English final test that constructed in the form of 40 multiple choice questions was firstly identified and grouped based on the classification of each domain of cognitive dimension according to Bloom's taxonomy revision that contained two levels namely HOTS (High Order Thinking Skills) and LOTS (Low Order Thinking Skills). Then, all students' answers were analyzed by using SPSS (Statistical Package for the Social Sciences) version 25 for windows to analyze the validity and reliability of the test. The result showed that the test questions just obtained 50% of applying (C3), 27.5% of understanding (C2), and 22.5% of remembering (C1) of LOTS levels whereas HOTS levels (analyzing, evaluating, and creating) weren't even found. Furthermore, its reliability quotient was just 0.729 (lower than 1.00) with 3 (8%) items categorized as invalid questions, and the valid items of 37 (92%) questions were predominantly classified in low interpretation. This research clearly showed that the test didn't fulfill the qualification of a good test.*

Key Words: *Validity, Reliability, Cognitive, Educational Taxonomies, Summative Test*

ANALISIS VALIDITAS DAN RELIABILITAS BUTIR SOAL TAKSONOMI KOGNITIF PADA TES SUMATIF BAHASA INGGRIS DI SMA NURUL FALAH PEKANBARU

Murni Wahyuni Sianturi, Dr. Erni, S.Pd., M. Hum, Dr. Dahnilsyah, SS., M. A.
Email: murni.wahyuni4690@student.unri.ac.id , erni@lecturer.unri.ac.id , danil_71@yahoo.com
Phone number: 081391661946

Mahasiswa Pendidikan Bahasa Inggris
Jurusan Bahasa dan Seni
Fakultas Keguruan dan Ilmu Pendidikan
Universitas Riau

Abstrak: Penelitian ini bertujuan untuk menganalisis kualitas tes sumatif yang disusun oleh guru di kelas XI SMA Nurul Falah Pekanbaru dan dikonstruksi dalam bentuk penelitian deskriptif kuantitatif. Sampel terdiri dari 53 siswa di semester kedua pada tahun pelajaran 2020/2021 dan dipilih menggunakan teknik total sampling serta metode undian. Selanjutnya, instrumen soal ujian semester Bahasa Inggris yang terdiri dari 40 butir soal pilihan berganda, terlebih dahulu diidentifikasi dan dikelompokkan berdasarkan masing-masing domain dimensi kognitif menurut revisi taksonomi Bloom yang memuat dua tingkatan yaitu HOTS (keterampilan berpikir tingkat tinggi) dan LOTS (keterampilan berpikir tingkat rendah). Kemudian, seluruh jawaban siswa dianalisis menggunakan SPSS (Statistical Package for the Social Sciences) versi 25 untuk windows dalam menganalisa validitas dan reliabilitas soalnya. Hasil menunjukkan bahwa soal tes hanya terdiri dari 50% tingkat menerapkan (C3), 27,5% memahami (C2) dan 22,5% mengingat pada tingkat keterampilan berpikir rendah, sedangkan keterampilan tingkat tinggi (menganalisis, mengevaluasi dan mencipta) bahkan tidak ditemukan. Selanjutnya, hasil bagi reliabilitasnya hanya 0,729 (lebih rendah dari 1,00) dengan 3 (8%) butir yang dikategorikan sebagai pertanyaan tidak valid dan 37(92%) butir yang valid sebagian besar diklasifikasikan dalam interpretasi rendah. Penelitian ini secara jelas menunjukkan bahwa tes tersebut tidak memenuhi kualifikasi tes yang baik.

Kata Kunci: Validitas, Reliabilitas, Kognitif, Taksonomi Pendidikan, Tes Sumatif

INTRODUCTION

Constructing a good summative test is important. According to Brown, H. D. (2010), "Test is a method of measuring a person's ability, knowledge, or performance in a given domain". Then, a summative test is a type of test that describes the whole parts of a lesson at the end of a long certain period for grading purposes, for example; the MID term test and final test. In other words, a summative test is an instrument or a tool that is needed to describe the students' ability at the end of a certain period in a given domain for grading purposes. Additionally, a good summative test shows the students' competence. It shows how students activate their critical thinking and analysis, solve given problems, and do the self-assessment from the beginning to the end of the lesson. It's also the type of an important test because it determines the final judgment. Furthermore, a teacher needs this test to see the effort of teaching instruction. A good teacher needs to examine the instruction, whether it's optimal or not, it's shown on the students' test result because students' ability is an abstract thing, but test result makes it concrete. Moreover, the summative test can be used to achieve the learning target or goal. A teacher needs to work hard on constructing this test because he or she has to ensure that it covers the whole part of the lesson from the beginning. Poor quality of a summative test is not more than just time-consuming if it doesn't show the students' real ability accurately. The type of test questions that are too easy or complicated cannot clearly describe the students' real competence. However, even constructing a good test is not a simple thing, but if the teacher constructs it well, the learning goal or target can be achieved optimally. In conclusion, constructing a good summative test is important because it helps the teacher to know the students' achievement, examine the teaching method or instruction, and to achieve the learning goal or target.

In addition, there are several tangles that mostly cause the poor quality of summative tests in SMA Nurul Falah Pekanbaru. Firstly, a teacher doesn't follow guidelines consistently. According to Lestari, M. W. 2021, (an English teacher of SMA Nurul Falah Pekanbaru) in an interview, following guidelines such as lesson plan or syllabus is complicated. She argues that sorting and constructing test questions based on the internet is preferable because it's easier than dealing with strict rules. However, the internet isn't a credible source because all ages can easily get access to it. Even students can instantly find the key answer to test questions so that test should be useless. Secondly, the teacher assesses students just based on her desire. A good test determines the students' real competence accurately. If a teacher wants to know the students' real competence, she has to consider the specified standard and do not assess students just based on her judgment. In other words, the teacher cannot construct a good test without dealing with rules and guidelines. Thirdly, the length of teaching cannot guarantee the quality of teachers. Lestari, M. W. (2021) states that she has experienced teaching as an English teacher for several years but still faces many problems on constructing tests. It means that the length of teaching cannot guarantee the teacher's mastery in assessment. Even a teacher knows all concepts or theories in the syllabus and curriculum, but if she doesn't apply it well, the result will be in contrast with the goal. For those reasons, the quality of a test isn't determined by the teacher's judgment or the length of teaching, but depending on how well and consistently a teacher follows guidelines and assesses students with the specified standard.

Besides, the implementation of Bloom's taxonomy in the summative test of SMA Nurul Falah Pekanbaru is still less than optimal. The taxonomy itself means an

educational framework that is used to classify the learning objectives into the level of complexity and specificity and to design a conceptualized curriculum and assessment. Lestari, M. W. (2021) in the interview, states that she assesses students and constructs tests by following Bloom's taxonomy revision and 2013 curriculum as a guideline. Benjamin Samuel Bloom (an educational psychologist) constructed this taxonomy firstly in 1956 and it was revised by Lorin W. Anderson (a student of Benjamin Samuel Bloom) and David R. Krathwohl (one of Benjamin Samuel Bloom's partners on designing the original taxonomy) in 2001. This revision has been translated into many languages and has provided a basis for test design and curriculum development throughout the world including Indonesia that implemented in the 2013 curriculum. Those who engage the educational pedagogics and implement the national 2013 curriculum in learning should pay more attention to this taxonomy because the 2013 curriculum itself is designed based on the level of human critical thinking skills of Bloom's taxonomy revision. According to Sulistyani (2019) in her research, she emphasizes that the 2013 curriculum has already been proposed by Nation and Macalister (a professor and a senior lecturer at Victoria University of Wellington, New Zealand who specializes in the fields of language teaching methodology and curriculum design and draws on experience in teacher education and curriculum design). Even though this curriculum has been qualified and claimed as an ideal character-building curriculum, on the other hand, the constraints faced by the teachers' lesson plans still do not refer to the 2013 curriculum (Gunawan, 2017). It means that teacher who deals with the 2013 curriculum will implement all arrangements and rules with consistency. In contrast, the lesson will never be properly taught without referring to the curriculum itself. According to Lestari, M. W. (2021), she states that the English summative test used in SMA Nurul Falah Pekanbaru is constructed only based on the LOTS levels. Moore and Stanley (2010) argue that there are two major groups of human thinking skills according to Bloom's taxonomy revision, they are; low-order thinking skills (LOTS), and higher-order thinking skills (HOTS). LOTS are the first three categories of learning domains taxonomy; remembering, understanding, and applying. However, HOTS are the last three categories of learning domains taxonomy; analyzing, evaluating, and creating. According to the learning purpose or target, both categories of HOTS and LOTS can be used to determine a good test, but based on the 2013 curriculum, HOTS is the main focus of the successful critical thinking assessment. In summary, the teacher has to pay attention to both LOTS and specially HOTS levels to design the optimal assessment of summative test that refers to the 2013 curriculum.

Above all, the teacher needs to know the characteristics of a good test. Mohajan (2017) states that there are two main characteristics of a good test, they are; reliable and valid. A test can be categorized as a reliable test when the result of the test can show the same or similar result after being tested a few times. Then, validity is the correlation between the result of the test and the students' real competence. A valid test can be used to represent the student's ability because both students' test results and the real abilities are related and similar. According to Bloom (1956), there are three domains in learning; affective, cognitive, and psychomotor. Through the test, teachers could describe the certain scale of learning outcomes in those three domains. Certainly, teachers have to construct fair and good tests to determine the successful study. Teachers need to consider the reliability and validity of measurement instruments because those criteria will determine the accuracy of the assessment. Without testing, it would be difficult for teachers to prove the learners' quality or to evaluate the learning process. Bloom also

explains, “The cognitive process begins from the simplest one, remembering what has been held, to the most difficult thing, deciding the value and worth of an idea”. It means that teachers have to deal with the rules, refer to the available syllabus, and couldn't construct a test just based on their desires. Those facts are the main reasons for the researcher to analyze the English summative test used in SMA Nurul Falah Pekanbaru.

Due to the importance of constructing a good summative test as a determiner of final judgment, this analysis research is conducted to show how the researcher examined the quality of the final test used by the eleventh-grade students in SMA Nurul Falah Pekanbaru by analyzing the level of cognitive domains (six domains) and criteria of a good test (validity and reliability) according to Bloom's taxonomy revision. The reason for choosing them as the population or participant of this research was that their syllabus refers to the 2013 curriculum. The researcher also consulted and confirmed with the English teacher that Bloom’s taxonomy had been applied as her guideline for constructing tests. Based on the 2013 curriculum, the HOTS must be implemented in the test. As it was described, the three components of HOTS were analyzing, evaluating, and creating. Then, to construct the test fairly, the teacher needs to make sure that the test is valid and reliable. In summary, the research is entitled “An Analysis of Validity and Reliability of Cognitive Taxonomy Items on English Summative Test at SMA Nurul Falah Pekanbaru”.

METHODOLOGY

This research was conducted at SMA Nurul Falah Pekanbaru as descriptive quantitative research. Sugiyono (2008: 13) defined that descriptive research as research to determine the result of independent variables without making comparisons or linking to other variables. According to Muijs (2004), quantitative research is collecting numerical data to explain a particular phenomenon and particular questions. Therefore, based on the title "An Analysis of Validity and Reliability of Cognitive Taxonomy Items on English Summative Test at SMA Nurul Falah Pekanbaru", this research had only one variable, namely test. It wasn't compared with another variable. Furthermore, this research analyzed the quality of tests constructed by a teacher by using Bloom's taxonomy.

The population of this research was the eleventh grade of SMA Nurul Falah Pekanbaru in the academic year 2020/2021. There were three grade classes of this school with the number of students was 116 as shown in table 1. They were selected by using total sampling and a lottery method to know which grade would be the sample. The researcher prepared three small pieces of paper. Each of those papers was written with the words 'tenth grade', 'eleventh grade', and 'twelfth grade'. Those pieces of paper were rolled and put in a box. Then, the English teacher was required to pick one of those pieces. The group of classes that were selected would be the sample of the research. After going through the sampling process, Class XI was chosen as the sample of the research whereas the sample consisted of 53 students.

Table 1. The Population of Students at SMA Nurul Falah Pekanbaru in academic year 2020/2021

Name of Class	Number of Students
Class X	24
Class XI	53
Class XII	39
Total	**116**

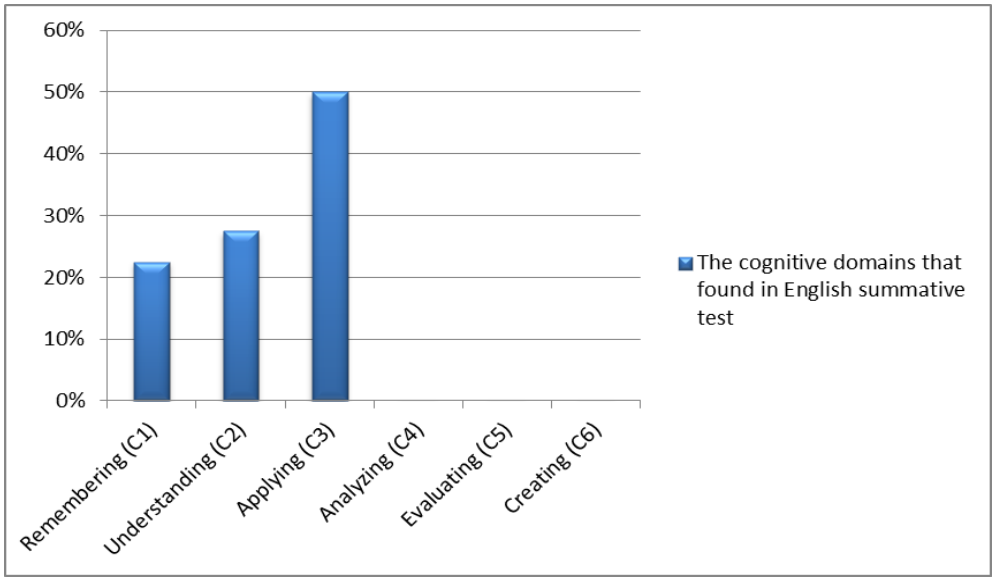
To collect the data, the writer used the teacher's syllabus, English final test questions, and students' answer sheets at the eleventh grade of SMA Nurul Falah Pekanbaru. She analyzed each cognitive level that was used according to the revision of Bloom's taxonomy. It was calculated in the form of percentages and presented in a diagram. The calculation of each percentage showed the LOTS and HOTS that were used to determine whether the English final test was appropriate with the syllabus and curriculum or not. Therefore, the writer calculated the content validity and the internal consistency reliability by using SPSS software. Those calculations showed the quality of English final test questions and presented in table and figure.

RESULTS AND DISCUSSIONS

The Result of the Cognitive Domains Analysis

After the data instrument analysis, the results of this research showed that the English final test questions used by the eleventh grade of SMA Nurul Falah Pekanbaru just obtained the three components of LOTS; 50% of applying (C3), 27.5% of understanding (C2) and 22.5% of remembering level (C1). According to the 2013 curriculum, HOTS are the main components that must be applied in assessment. However, the percentage of analyzing (C4), evaluating (C5), and creating (C6) were not found (0%) in this instrument, as shown in Figure 1.

Figure 1. The Diagram of Cognitive Level in the Final Test Used by the Eleventh Grade of SMA Nurul Falah Pekanbaru



The Result of Content Validity Analysis

Furthermore, the researcher analyzed and calculated the content validity by using SPSS software that set out in the formula according to Arikunto (2013). The description that shown was 37 (92%) valid questions and 3 (8%) invalid questions that were obtained and presented in figure 2. Therefore, to categorize the coefficient correlation between the score of the test and the score of the criterion the writer set the interpretation of validity in figure 3.

Figure 2. The Diagram of Content Validity Percentages in English Final Test Used by the Eleventh Grade of SMA Nurul Falah Pekanbaru

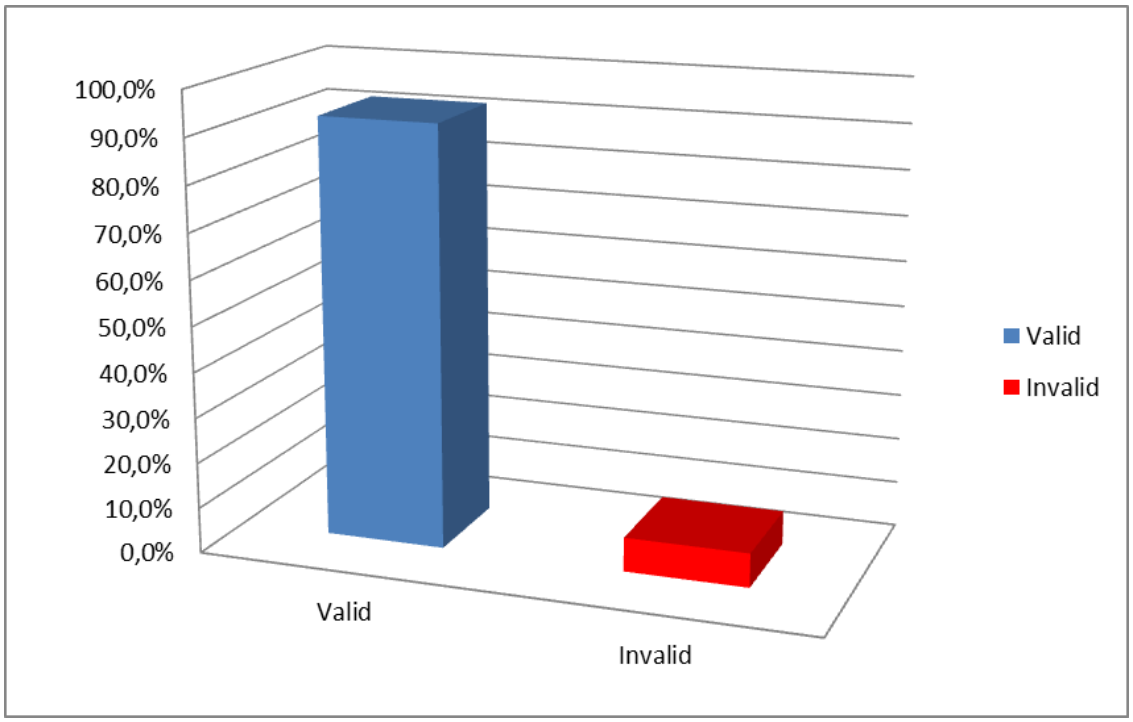
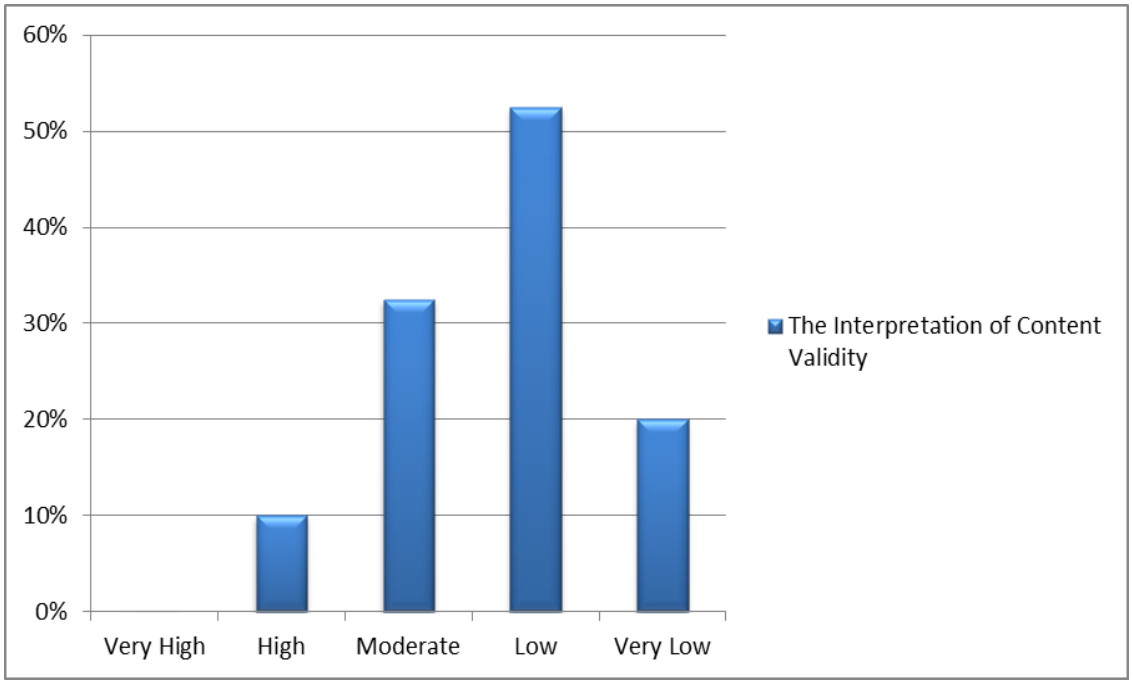


Figure 3. The Diagram Interpretation Validity in English Final Test Used by the Eleventh Grade of SMA Nurul Falah Pekanbaru



The Results of the Cognitive Domains

Moreover, the writer analyzed the reliability of the test also by using SPSS software. A reliability coefficient (r11) of 1.00 would indicate that a test is perfectly reliable. However, the test could not be categorized as reliable if the result showed less than 1.00 of reliability coefficient (r11). From the overall calculation, the English final test questions made by the teacher have 0.729 reliability index, as shown in figure 4. The reliability index was lower than 1.00.

Figure 4. The Interpretation of Reliability by using SPSS

Reliability Statistics

Cronbach's Alpha	N of Items
.729	41

Discussions

The findings of the research showed the classification of revised cognitive taxonomy level and the quality of a good test. The instrument of English final test questions used by the eleventh grade of SMA Nurul Falah Pekanbaru just obtained the three components of LOTS; 50% of applying (C3), 27.5% of understanding (C2), and 22.5% of remembering level (C1). However, according to the 2013 curriculum, HOTS are the main components that must be applied in assessment. The percentage of analyzing (C4), evaluating (C5), and creating (C6) were not found (0%) in this instrument.

Furthermore, the English final test questions used at SMA Nurul Falah Pekanbaru cannot be perfectly categorized as a reliable test because its reliability quotient is below 1.00. There were also still 3 (8%) items which categorized as invalid questions where the valid items of 37 (92%) questions were even predominantly classified in low interpretation. The ability of the teacher on constructing tests was categorized as the average to the low level. In other words, the English final test questions used by the eleventh grade of SMA Nurul Falah Pekanbaru didn’t refer to the 2013 curriculum and weren't suitable to be used as an ideal summative test for grading purposes.

CONCLUSION AND RECOMMENDATION

Conclusion

After analyzing the English final test questions made by the English teacher of SMA Nurul Falah Pekanbaru, it can be concluded that the test cannot be categorized as a good summative test. Based on the 2013 curriculum, HOTS are the main focus of the successful critical thinking assessment and a good test must be valid and reliable. Unfortunately, the English final test questions that made by the teacher didn't allow students to express their judgment, opinion or evaluation and were just consisted of reading and grammar test where the HOTS components (analyzing (C4), evaluating (C5), and creating (C6)) weren't found but the LOTS components (remembering (C1), understanding (C2) and applying (C3)) were obtained. The researcher recommends the English teacher combine both multiple choices and essays in the test because the essay test allows the teacher to assess the HOTS aspect and allows students to give their judgment, opinion, or evaluation. The researcher also recommends the English teacher have a better understanding of six taxonomy levels from C1 to C6 so that the test should be appropriate with those components according to the syllabus that refers to the 2013 curriculum. Moreover, the test cannot be perfectly categorized as a reliable test and still needs revision in several invalid questions because even the valid items of interpretation validity that were found just mostly categorized in the moderate to the low level and not in the high level. Finally, regarding the importance of validity and reliability, the researcher suggests the English teacher should study more about validity and reliability as the criteria of a good test.

Recommendation

The result of this research is expected to give some valuable recommendations theoretically and practically to the following people:

1. Theoretically, the researcher recommends the teacher to construct the summative test by following the standard of good test that refers to the 2013 curriculum.

2. Practically, it is expected to help other researchers who are interested in doing research related to this similar theme or other research that may be relevant to Bloom's taxonomy.

3. This research is also expected to give basic information to enhance the reader's understanding and knowledge about the level of cognitive domains (six domains) and criteria of the good test (validity and reliability) that is used for constructing English summative tests.

BIBLIOGRAPHY

Anderson, L. W., & Bloom, B. S. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York*: Longman.

Brown, H. D., & Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices* (Vol. 10). White Plains, NY: Pearson Education.

Gunawan, I. (2017). *Indonesian Curriculum 2013: Instructional Management, Obstacles Faced by Teachers in Implementation and the Way Forward*. In 3rd International Conference on Education and Training (ICET 2017) (pp. 56-63): Atlantis Press.

Mohajan, H. (2017). *Two Criteria for Good Measurements in Research: Validity and Reliability*. Chittagong: Munich Personal RePEc Archive (MPRA).

Muijs, D. (2004). *Doing Quantitative Research in Education*. London: Sage Publication Ltd.

Sugiyono. (2008). *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif dan R&D*. Bandung: Alfabeta, 13.

Sulistyani, U. N. L. (2019). *Level of Knowledge Construction in Elementary School Students: Lesson Plan Analysis*. Jurnal Pendidikan Humaniora, 7(4), 157-165.