English Education &
Applied Linguistics
Journal

THE ANALYSIS OF THE QUALITY OF NATIONAL STANDARDIZED SCHOOL SUMMATIVE EXAMINATION OF ENGLISH SUBJECT 2019

Nilam Ayuning Herdianti[1], Rajji K. Adiredja[2], Asep Suparman[3]

Institut Pendidikan Indonesia,
Garut, Indonesia[1,2,3]

Email:
asep.suparman@institutpendidikan.ac.id[1]

Article History
Submitted 23 June 2021
Revised 6 July 2021
Published 30 August 2021

Abstract
The national standardized school summative examination (USBN) is the test for measuring student competency achievement, it is constructed by Ministry of Education and Culture (Kemendikbud) about 25% and MGMP Province 75%. However, in improving the quality of the test, it is addition the cognitive dimention of High Order Thinking Skills (HOTS) as the standard. The purpose of this study is to prove the quality of the test in terms of item analysis, distribution of hots and consistency of the test. This study was conducted by analyzing the national standardized school summative examination (USBN) 2019 items, answer sheet, key answer and syllabus.The data was collected from archieves in one of Senior High School. The result shows that the item analysis of the test is valid and reliable, the level of difficulty is not balanced, the discriminating power is sufficient, and the effectiveness of distractor is mostly need revision. Moreover, the distribution of hots has pass the limit and the test consistency is satisfactory in criteria of a good test.

Keywords: National Standardized School Summative Examination, Item Analysis, High Order Thinking Skills

INTRODUCTION

The national standardized school summative examination (USBN) is the test for measuring student competency achievement carried out by the Education Unit by referring to Graduates' Competency Standards to gain recognition for learning achievement (Juknis USBN, 2018). It is constructed by Ministry of Education and Culture (Kemendikbud) about 25% and MGMP Province 75%. MGMP Province is groups of similar subject teachers at the Regency/City level at the Senior High School level. It has an important role in education and it has been called the greatest single social contribution of modern psychology, it may be the most useful evaluation method available for human resource intensive endeavours (Phelps, 2008; Tosuncuoglu, 2018). Hereafter, the cognitive dimension of high order thinking skills has been used as the standard for examination to improving the students critical thinking, it requires a great cooperation between all teachers of different subjects, the thoughtful consideration of current instructional techniques and the commitment to an active student-centered learning environment in different levels of studying to work together to achieve that goal (Limbach & Waugh, 2010; Abosalem, 2016). According to Permendikbud No. 81A of 2013 about the curriculum implementation stated that future competence is needed by individual who has high order thinking skills critically, communication skills, and creative. Moreover, there is consensus that 21st century education should prioritize students' skills for higher order thinking, transfer, and flexible reasoning over memorization of disciplinary facts (Richland & Simms, 2015). In line with Suprayitno in Maulipaksi (2019) that the composition of the questions is divided by cognitive level, which is 10-15 percent for reasoning or higher order thinking skills (HOTS). Therefore, because this test is designed to measure the student`s achievement competencies and determine student`s graduation from the school, it is very significant to conduct test items evaluation since it gives a clear portrait of the quality of the items and the test as a whole (Narwianta, Bharati, & Rukmini, 2019).

In the previous study, there are also some other studies on the national standardized school summative examination (USBN). Nurfiqah et al, (2015) found that the test items have given a big contribution for the teacher to measure the students' competence. Afterwards, it is indicated that questions asking low order thinking skills still prevailed in the test items and also showed the complete absence of "Appreciation" – the highest level of thinking in the mentioned taxonomy on the items (Ahmad, 2016). It is similar with Wasis et al, (2017) that the highest percentage of the test still measures the cognitive process of low order thinking skills if it is compared with PISA and TIMSS items. Moreover, the test produce direct educational benefits for students (Benjamin & Pashler, 2015). Consequently, the quality of test should be proven. It can using item analysis which can provides valuable information to the teachers to further item modification and future test development and offer educational tools to assist them (Siri & Freddano, 2011). Furthermore, it is essential in improving items which will be used again in later tests; it can also be used to eliminate misleading items in a test (Quaigrain & Arhin, 2017).

Instead of studying the advantages of test and its implication for students, the researcher is interested in proving the quality of national standardized school summative examination (USBN) through the consistency of the high order thinking skills between the instructional design and item analysis. Moreover, to know the distribution of each type of high order thinking skills used on the test. It is because teachers are supposed to implement varieties of assessment methods and stay away from the tests that require recalling knowledge (Doganay & Bal, 2010). Thus, the difference in this research with previous research is to prove the quality of national standardized school summative examination (USBN) 2019 of English subject. As a continuation of the results from the previous learning, so that will show how far the test indicate the extent of the high order thinking skills that have been applied. It can reveal the student's weakness and strength areas; the strength area to be enhanced and the weakness area to be treated (Abosalem, 2016).

METHODOLOGY

This research is a documentary study. In order to analyze the quality, the researcher used descriptive qualitative method. The quality refers to item analysis, the distribution of high order thinking skills on the test items and the consistency of the test. Item analysis includes item validity, item reliability, level of difficulty, discriminating power and distractor (Arikunto, 2018) which obtained by analyzing students answer sheet and key answer. Furthermore, in analyzing the distribution of high order thinking skills on the test items the researcher used Krathwohl and Anderson (2001) theory about a taxonomy assessing from a revision of

Bloom's taxonomy and The HOTS assessment rubric adapted from Directorate of High School Development Directorate General of Primary and Secondary Education (2017). Then, the syllabus used to analyzed the consistency of the items. In addition, the document analysis used in this research because it is widely applied for written or visual data with the purpose of identifying specific, characteristic of materials that are going to be analyzed in general form of textbook, newspaper or any other host of documents (Donald Ary et al, 2010).

The data of this present study were a collection of items archives in one of senior high school, i.e national standardized school summative examination or USBN 2019 which have been tested to students in twelfth grade, students answer sheets, key answer and syllabus. The test items contain 45 items include listening and reading comprehension. The items of listening comprehension are 10 items and the rest are for reading comprehension. The items consist the material from first grade until the third grade. The researcher got the data on May 2019.

In order to analyze the item analysis refers to item validity, item reliability, level of difficulty, discriminating power and distractor the researcher was helped by Anates 4.0 and Microsoft Excel. Then, in analyzing the HOTS type on the items the researcher used Bloom’s cognitive taxonomy revised by Anderson and Krathwohl (2001) and HOTS assessment rubric adapted from *MODUL pembuatan soal HOTS* (2017). Consistency of the items was analyzed by matching the material of each item with the material in syllabus

FINDINGS AND DISCUSSION

1. Findings a. Item Analysis

In this study there were several points that have been analyzed by the researcher, including item validity, item reliability, level of difficulty, discriminating power and distractor (Arikunto, 2018).

Table 1. Item Validity, Item Reliability, Level of Difficulty and Discriminating Power

Validity			Reliability	Level of difficulty				
Valid	Invalid			Very Easy	Easy	Moderate	Difficult	Very Difficult
35 items	5 items		0,74	11 items	8 items	14 items	5 items	2 items

	Discriminating power			
Excellent	Good	Satisfactory	Poor	Very Poor
1 item	9 items	16 items	10 items	4 items

From the table above, it shows that there are 5 items from 40 items which is invalid. The reliability of the items can be declared reliable on a high scale because it has a range of values between 0,60 to 0,80 as quoted from the Guilford criteria (Sundayana, 2018). Moreover, the assumption used to obtain good quality items besides fulfilling validity and reliability is a balance of the difficulty level of the items. The balance that refers to the existence of items which are proportionally very easy, easy, moderate, difficult and very difficult. There are 11 items which categorized as very easy items, 8 easy items, 14 moderate items, 5 difficult items, and 2 very difficult items. In addition, the analysis of the discriminating power conducted on the whole items known that there are several items that have very poor discriminating power. From the 40 multiple choice items that were tested, items with excellent discriminating power were only 1 item (2,5%), items with good discriminating power were 9 items (22,5%), items with satisfactory discriminating power were 16 items (40%), items with poor discriminating power were 10 items (25%), and 4 items (10%) with very poor discriminating power. Based on the results of this study it can be concluded that 65% of multiple choices items have sufficient discriminating power and 35% of items have weak discriminating power.

Table 2. Items Distractor Category

No	Distractor Category	Items Number	Total
1	Accepted	2, 5, 10, 11, 13, 17, 20, 21, 23, 24, 25, 28	12
2	Rejected	1, 33, 40	3
3	Repaired	3, 4, 6, 7, 8, 9, 14, 15, 16, 26, 29, 31, 34, 38, 39	15
4	Can be rejected or repaired	12, 18, 19, 22, 27, 30, 32, 35, 36, 37	10

The effectiveness of using distractor can be known by looking at the pattern of the distribution of students answer. The pattern of answer distribution is obtained by counting the number of students who chose answers a, b, c, d, e, or did not choose any answer. Based on the pattern of answer distribution, it can be determined whether the distractor function is good or not. Distractor works well when alternative answers are chosen for at least 5% of all test takers. The number of subjects in this study was 476 students, so the distractor would function well if at least 5% of 476 students were selected, namely 24 students.

The result from the table shows the items distractor number 1, 33 and 40 mostly have bad distractor. Items number 2, 23 and 25 have 2 sufficient distractor and 2 very good distractor. Items number 3, 7, 9, 29 and 38 have one of each distractor of bad, sufficient, good and very good. Item number 4 has 1 very bad, 1 sufficient and 2 good distractor. Item number 5 has 1 sufficient, 1 good and 2 very good distractor. Items number 6, 8, and 26 have 1 bad, 2 sufficient and 1 good distractor. Item number 10 has 3 sufficient and 1 good distractor. Item number 11 have 1 sufficient and 3 very good distractor. Items number 12, 18 and 19 have 1 very bad, 1 bad and 2 sufficient distractor. Item number 13 has all good distractor. Items number 14 has 1 very bad and 3 sufficient distractor. Item number number 16 and 31 have one of each distractor of very bad, sufficient, good, and very good. Item number 17 has 1 sufficient, 2 good and 1 very good distractor. Items number 20, 21, 24, and 28 have 2 sufficient and 2 good distractor. Items number 22, 30 and 35 have one of each distractor of very bad, bad, good and very good. Items number 27 and 36 have 2 bad and 2 sufficient distractor. Item number 32 has 2 bad, 1 good, and 1 very good distractor. Item number 34 has 1 very good, 2 sufficient, and 1 good distractor. Item number 37 has one of each distractor of very bad, bad, sufficient, and good. Item number 39 has 1 bad, 1 good and 2 very good distractor. The last item number 40 has 1 very bad, 2 bad and 1 sufficient distractor.

b. The Distribution of High Order Thinking Skills

This section explained the analysis of the items by applying HOTS assessment rubric. There were 45 items which were analyzed, the result shows in the following below;

There are about 17 items (39%) which distributed as high order thinking skills and 27 items (61%) as low order thinking skills.

Table 3. The Distribution of High Order Thinking Skills

Cognitive Taxonomy	I. Remembering	II. Understanding	III. Applying	IV. Analyzing	V. Evaluating	VI. Creating
Items	11, 15, 22, 25, 26, 32, 40, 44	5, 6, 8, 10, 14, 18, 20, 23, 27, 30, 31, 37, 39, 45	1, 2, 3, 4, 34	13, 16, 17, 19, 21, 24, 29, 33, 38	7, 9, 12, 28, 35, 36	41, 43

The items which categorized as C1 (Remembering) are question that asking particular information which available on the passage. Items which included C2 (Understanding) are mostly question about stating main idea of the passage, however the rest of the question is the synonym of word, interpreting the audio to the picture. In cognitive taxonomy C3 (Applying) are question to applying the best answer to the monolog and blank sentence. Moreover, in C4 (Analyzing) the items are about to distinguishing the relevant from irrelevant parts of the text and making inference by interpreting the writer intention from the text. Then, in C5 (Evaluating) the question is about stating readers opinion based on information. In addition, C6 (Creating) is the question about arranging the text and making caption from the chart.

c. The Consistency of Items

Based on its contents validity there are 17 material from the 1st until the 3rd grades basic competences which exist on the test, so there are about 5 basic competences material is not used on this test. The items that consist basic competences material from the 1st grades are about 14 items, the 2nd grades about 13 items, and the 3rd grades about 17 items. The material from the 1st grades is about 5 material, the 2nd grades about 6 material and the 3rd grades about 6 material.

Through the analysis were conducted, there are several items that invalid in content validity and construct validity, only content validity, and only construct validity. There are about 2 items that invalid in content validity and construct validity, 10 items that invalid in content validity, and 1 item that invalid in construct validity. The items detail is on the following table;

Table 4. The Category of Invalid Items Content and Construct Validity

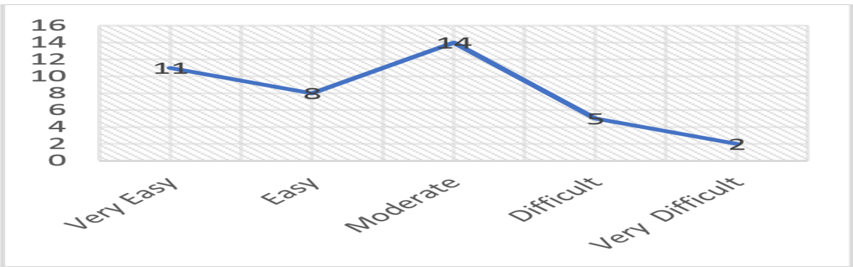
Category	Items Number

Invalid Content Validity and Construct Validity	26, 35
Invalid Content Validity	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Invalid Construct Validity	45

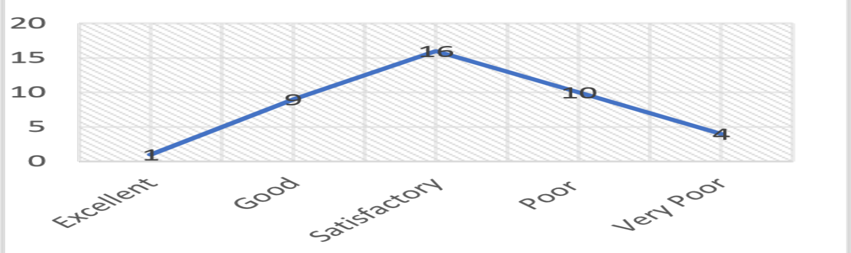
DISCUSSION

In this study, the national standardized school summative examination (USBN) 2019 has used as data collection. This discussion was intended to describe the quality of the test.

There are three aspect of quality that have analyzed by the researchers; the item analysis, distribution of HOTS, and consistency of the test. Based on the previous result in item analysis, the validity and reliability of the items is highly valid and reliable with items that 88% valid and 74% reliable. The items difficulty and the discriminating power of the items in the following curve.



Curve 2. The Discriminating Power of The Items



Curve 1. The Items Difficulty

As it shown in the curve, the items difficulty of the test is not balanced. According to EXHCOBA in Escudero, Reyna, and Morales (2000) that the median difficulty of the items should range between very easy 5%; easy 20%; medium 50%; difficult 20%; very difficult 5%. Hereafter, although there are about 35% items that indicated have low discriminating power, but about 65% items have decent discriminating power ability. It can be claimed that the items suitable with Kusnandar (2014) that the multiple choice test requirements sufficient discriminating power to distinguish high achieving students from low-achieving students.

The result shown that the accepted distractor only have percentage about 30%. The rest about 70% should be rejected (7%), repaired (38%), and rejected or repaired (25%). In line with Sudijono (2009) that the distractor that has been able to carry out its functions properly can be used again in future tests, while the distractor that has not been able to function properly should be repaired or replaced with another distractor.

Subsequently, after investigating the HOTS distribution on the test, it has found that the test mostly still on low order thinking skills (61%). Whereas, the National Examination in 2018 results indicate that students are still weak in HOTS (Setiawati, Asmira, Ariyan, Bestary, & Pudjiastuti, 2018). However, according to Totok Suprayitno in Maulipaksi (2019) that the composition of the questions is divided by cognitive level, which is 10-15 percent for reasoning or higher order thinking skills (HOTS). It means that the test has exceeded the limit that should be achieved with a number which is not small at 39%. This improvement is a good thing, considering that the test must be upgraded to HOTS quality. In line with Country Note – Results from PISA 2015, (2015) if students can keep up that pace of improvement, they will have a realistic chance to match the science performance of their peers in the industrialised world by 2030, the year for which the United Nation's Sustainable Development Goals expect every student to benefit from quality education.

Lastly, the consistency of the test can be seen from the good test criteria; practically, reliability and validity (Douglas, 2000). Practically, this test is not wasting times because only held in 90 minutes, it ease on scoring because the multiple choice test is rated by machine, teachers only have to scoring on the essay items, annd financially this test does not make students to spend money. The reliability of this items are checked before in items analysis. Regarding with its validity there are three types of validity, contents validity, construct validity and face validity. After observing the result of the data analysis there is about 30% items that should be revised because invalid in content validity and construct validity, only content validity and only construct validity. Yet, the 70% are in accordance with basic competencies and question indicators this is similar with Douglas (2000) and Hughes (2003) that it is represent sample of the language skill. The items which invalid in content validity and construct validity are the reading comprehension. Whilst invalid content validity are all of items from listening section. Only one item that failed on construct validity.

According to (Arikunto, 2018) the test for each topic should balance and unambiguous. From the test the researcher found that there is misconception from the test; the listening section has different content with the option, there are one basic competence that have 4 items on the test, even though there are five basic competence that not included to the test. In addition, one items have ambiguous answer.

Finally, in term of face validity the test is valid. Concerning with the part of listening section and reading comprehension which is not testing another skill. It is reflect what it should be testing. As Hughes (2003) that it is said to have face validity if it looks as if it measure what it is supposed to measure.

CONCLUSION

This study identified the quality of national standardized school summative examination (USBN) of English subject in terms of item analysis, the distribution of high order thinking skills on the test items and the consistency of the test. The result obtained from the data analysis in terms of item analysis of the validity and reliability is depending on the value of test, the more higher value the more valid and reliable a test. The level of difficulty of this test only balanced in easy and very difficult category which means the test yet fulfil the proportional of items difficulty. Hereafter, the discriminating power of this test is sufficient to distinguish high-achieving from low achieving students. Then, items distractor mostly should be repaired or replaced in order to function. Moreover, the distribution of HOTS in this test is high compared to its limit, this progress can be appreciated and maintained for further tests. Hence, it will increase the chance in improving the education quality from assessment field.

Furthermore, the test consistency is satisfactory in criteria of a good test, it is practically in time, financial and administration. Regarding with its validity the content validity and construct validity is mostly valid. The invalid test are in items that miss conception and ambiguous. Thus, the face validity is valid cause the test measures what it is supposed to measure.

REFERENCES

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

Ahmad, U. L. (2016). Senior high school english national examination and thinking skills. *Beyond Words*, 168188.

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for learning, teaching and assessing, a revision of Bloom's taxonomy of educational objectives.* (A. Martinez-Ramos, Ed.) New York: Addison Wesley Longman, Inc.

Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan.* Jakarta: Bumi Aksara.

Benjamin, A. S., & Pashler, H. (2015). The value of standardized testing: A perspective from cognitive psychology. *Policy Insights from the Behavioral and Brain Sciences 2(1), 2,* 13-23. doi:10.1177/2372732215601116

Country Note – Results from PISA 2015. (2015). *Programme for International Student Assessment (PISA)*, 1-8.

Doganay, A., & Bal, A. P. (2010). The measurement of students' achievement in teaching primary school fifth year mathematic classes. *EDUCATIONAL SCIENCES: THEORY & PRACTICE*, 200-215.

Douglas, H. B. (2000). *Teaching by principles : An interactive approach to language pedagogy* (2nd ed.). San Francisco, California: Longman.

Escudero, E. B., Reyna, N. L., & Morales, M. R. (2000). The level of difficulty and discrimination power of the basic knowledge and skills examination (EXHCOBA). *Revista Electrónica de Investigación Educativa, 2*, 4.

Retrieved from http://redie.uabc.mx/vol2no1/conte nts-backhoff.html

Hughes, A. (2003). *Testing for language teachers.* Cambridge: CAMBRIDGE UNIVERSITY PRESS.

Limbach, B., & Waugh, W. (2010). Developing higher order thinking. *Journal of International Pedagogies*, 1-9.

Maulipaksi, D. (2019, March 27). *Tingkat komposisi soal UN tidak berubah, Ini komposisi soalnya*. Retrieved from www.kemendikbud.go.id: https://www.kemdikbud.go.id/main /blog/2019/03/tingkat-kesulitansoal-un-2019-tidak-berubah-inikomposisi-soalnya

Narwianta, N., Bharati, D. A., & Rukmini, D.

(2019, September 15). The evaluation of higher order thinking skills in english school nationally standardized examination at state senior high school 6 semarang. *English Education Journal*.

Nurfiqah, S., Supardi, I., & Novita, D. (2015). The analysis on the items of the english test made by the teacher . *Jurnal Pendidikan dan Pembelajaran Khatulistiwa*, 1-10.

Phelps, R. P. (2008, June 3). The role and importance of standardized testing in the world of teaching and training. *Paper presented at the 15th Congress of the World Association for Educational Research*, 1-9.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Educational Assessment and Evaluation Research Article*, 1-11.

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking,. 1-16. doi:10.1002/wcs.1336

Setiawati, W., Asmira, O., Ariyan, Y., Bestary, R., & Pudjiastuti, A. (2018). *Buku penilaian berorientasi high order thinking skills.* Jakarta: Direktorat Jenderal Guru dan Tenaga Kependidikan.

Siri, A., & Freddano, M. (2011). The use of item analysis for the improvement of objective examinations. *International Conference on Education and Educational Psychology* , 188-197.

Sudijono, A. (2009). *Pengantar evaluasi pendidikan.* Jakarta: Rajawali Pers.

Sundayana, R. (2018). *Statistika penelitian pendidikan.* Bandung: Alfabeta.

Tosuncuoglu, I. (2018, September). Importance of assessment in ELT. *Journal of Education and Training Studies,* *6*(9), 163-167 doi:10.11114/jets.v6i19_3443-