Senior High School English National Examination and Thinking Skills

Ummu Lathifah Ahmad
ummulathifah.ahmad@yahoo.com
East Java Provincial Education Department
Surabaya, Indonesia

Abstract

When English National Examination (abbreviated into ENE) as a norm-referenced test is designed for instructional purposes, to evaluate the result of national curriculum, it is very significant to conduct item test evaluation since it gives a clear portrait of the quality of the items and of the test as a whole. The purpose of this study was to analyze which levels of the Barrett taxonomy were more reflected in ENE items of 2013/2014 academic year and whether the proportions of items among the twenty test packages in the ENE assessing students' Lower Order Thinking Skills (LOTS) and Higher Order Thinking Skills (HOTS) are consistent. The researcher adopted the qualitative descriptive approach using a content analysis card to codify the ENE items. To ensure the reliability of the study, three inter-raters analyzed a sample of the test packages. The results indicated that questions asking LOTS still prevailed in ENE items. Of all the twenty test packages, the items categorized into literal level represented around 68.6% of the total number of the questions. Meanwhile, the questions belonging to reorganization came to occupy a percentage of 20.8 and the questions asking the students' inferential level only reached 10.3%. Also, the tests were not enriched sufficiently with the evaluation comprehension since they only comprised 0.3%. The results also showed the complete absence of "Appreciation" – the highest level of thinking in the mentioned taxonomy. It is obvious that there is a shortage of items questioning students' HOTS in the exam and they are not well-treated. Accordingly, this finding reveals that there is still much room for ENE to be the driving force in the effort to make learners critical thinkers. In the light of these data, this study recommends modifying the English National Exam by providing them with more question items that include HOTS.

Keywords: Content analysis, Barrett's Taxonomy, English National Examination

Introduction

Measuring students' proficiency in particular skills of the language requires teachers and others in evaluative positions to develop a systematic procedure of language testing. A language test can be of any scale to gauge some qualities of students after participating in learning a particular language for some period. Besides, it can be a precious tool for obtaining information that is relevant to several concerns in language teaching, such as providing evidence of the results of learning and instruction which in turn serve as feedback on the effectiveness of

the teaching program itself, providing information that is relevant to making decisions about individuals, i.e. determining what specific kinds of learning materials and activities should be given to students (Bachman & Palmer, 1996, p. 8).

One type of tests is standardized test. Brown defines a good standardized test as the typical norm-reference test which aims to place test-takers "on the continuum across a range of scores" and to classify test-takers by their rank (2004). Standardized test is employed to measure the students' mastery on basic parts of the curriculum in general and the result functions as a portrait of our education quality.

An example of a large-scaled standardized test administered in Indonesia is the National Examination (abbreviated into NE) held annually throughout the country to measure students' achievement at the end of a learning period in each level. It is the latest form of a school leaving examination in Indonesia starting from 2005 until now. NE can be defined as a test to measure and evaluate the students' competence nationally by the central government after the process of teaching and learning (The Regulation of the Minister of Education 2005, p.1). It is implemented as a way of improving national education quality.

When NE as a norm-referenced test is designed for instructional purposes to evaluate the result of national curriculum, it is very important to conduct item test evaluation. The result can give a clear portrait of the quality of the items and of the test as a whole and can also be used to improve both items and the tests as a whole. Brown and Rodgers (2002, p. 289) define evaluation as "the process of seeking to establish the value of something for some purpose". To achieve this, evaluative processes on different fields of curriculum ranging from learning, teaching and assessing should be carried out to find out the strengths and weaknesses as well.

Good test items are those items that can assess the performance of learners effectively. Since language testing has such a powerful influence on classroom instruction, it is important for educators to be informed about the question types in examination, especially a high-stake exam such as the National Exam. With this knowledge, educators can evaluate the level of comprehension and the students' competence to process high order thinking skills. Students' interactions with questions directly influence their future learning outcomes (Armbruster & Ostertag, 1993). The implication is that higher order questions would promote higher order processing of the text.

This study is primarily anchored on the Barrett's Taxonomy of Comprehension, which discusses the different levels of Comprehension namely: literal, reorganization, inferential, evaluation and appreciation. The theory assumes that learners move from the literal understanding to another, until the learner fully understands and appreciates the cognitive and aesthetic aspects of the material. The first two categories, literal and reorganization comprehension, deal with the facts as presented orally or in the books the students have read, and thus result in closed questions that have a single correct response. Inferential comprehension is demonstrated when students use the ideas and information explicitly stated in a viewing material, students' intuition and personal experiences as bases in making intelligent guesses and hypothesis. Evaluation comprehension refers to judging the language and effect of the material in the light of appropriate criteria. It requires responses which indicate that an evaluative judgment has been made by comparing ideas. Appreciation comprehension deals with psychological and aesthetic responses. It refers to emotional responses to content, plot or theme, sensitivity to various literary genres, identification with characters and incidents, reaction to author's use of language, and response to generated images. The remaining categories always involve the student's own background knowledge. Consequently, many different, but correct, responses will emerge since each student owns a different background of home, family, friends, and learning process. These categories therefore lead to the development of open-ended questions which require students to use higher order thinking skills.

One interesting aspect of the Barrett taxonomy, according to Armbruster & Ostertag (1993), is the subdivision of categories according to specific type of information targeted by the question (e.g. recognizing and recalling main ideas, inferring cause and effect relationships, identification with characters and incidents). It contributes to the usefulness of Barrett's taxonomy as a guide for constructing questions on a variety of levels as well as for judging questions that have already been created. It can be used to evaluate students' comprehension of text. Bloom's taxonomy of higher thinking skills sheds light on Barrett's comprehension as illustrated in Table 1[1].

The right column displays two categories according to the required level of cognitive operation: Lower-Order Thinking Skills and Higher-Order Thinking Skills. The first demands the

[1] Table 1, p.17

recognition or recall of factual information explicitly presented in the text. The information generally involves facts, names, dates, times, locations, lexical items, and propositions. Literal comprehension and reorganization fall into Lower-Order Thinking Skills category since questions of literal comprehension and reorganization can be answered directly and explicitly from the text. On the other hand, Higher-Order Thinking Skills require more than mere recognition or recalling information. They also facilitate moving beyond a literal understanding of the text to a more knowledge-based and global understanding of textual meaning. In other words, they require readers to read beyond the lines. Thus, inferential comprehension, evaluation and appreciation belong to Higher-Order Thinking Skills because in order to answer these types of question, students must use both a literal understanding of the text and their knowledge of the text's topic and related issues.

Researchers have shown that comprehension skills and success in learning L1 and L2 as well as other subjects are closely related. Thus, the comprehension skills should be taught to train students' cognitive skills ranging from literal comprehension to appreciation comprehension. When these skills are practiced, students can develop not only their lower order thinking skills (LOTS) but also their higher order thinking skills (HOTS) and can effectively respond to testing items which assess the latter skills. LOTS is the foundation of skills required to move into higher order thinking. These are basic skills that are taught very well in school systems and include activities in reading and writing (Wilson, 2000).

Due to this fact, it can be argued that HOTS are fundamental skills that can empower individuals' ability to analyze, to synthesize (to combine knowledge of different sources), to discuss, to judge, and to evaluate (McDavitt: 1993, p. 20). It is also in line with Grigaite's findings (2005), who investigated the effect of using higher order thinking strategies on developing child's thinking skills. Fifty-seven children at the age of six took part in the research. Findings revealed that students in the experimental group who participated in the training were creative. They further revealed high degrees of cognitivism. In addition, Tomei (2005) defines "HOTS involve the transformation of information and ideas. This trans-formation occurs when students analyze, combine facts and ideas and synthesize, generalize, explain, or arrive of some conclusion or interpretation. Manipulating inform-ation and ideas through these processes allows students to solve problems, gain understanding and discover new meaning."

It is worth noting that higher levels of thinking happens when learners "search beyond the content they are reading, to find out the answer or achieve comprehension" (Razmjoo & Madani, 2013). Predicting, concluding, inferring are instances of reading comprehension strategies that evoke higher levels of thinking. The level of items developed based on the taxonomy affects the performance of learners in answering reading comprehension items. What is more, it can be understood that a relationship exists between the level of thinking procedures required and the learners' ability to answer the item properly. The effects of using (HOTS) strategies do not only improve the learner's listening and reading comprehension, but also their thinking, brainstorming and writing abilities.

However, despite the significance of evoking students' higher order thinking skills, many test items are still designed to test students' LOTS. The reading comprehension questions mainly consist of literal and reorganization level which students can easily answer directly and explicitly from the text. As a result, students do not get accustomed to read beyond the lines, which require them to combine both a literal understanding of the text and their schemata. For instance, in their study, Razmjoo & Madani (2013) analyzed University Entrance Exam (UEE) items, in terms of Bloom's revised taxonomy, to find out which levels of this taxonomy were more reflected in these items. The results indicated that Lower Order Thinking Skills (LOTS) were more considered in UEE items. The findings also showed the complete absence of "Creating" which is the highest level of thinking in the mentioned taxonomy.

Another study was conducted by Humos (2012) who analyzed reading comprehension questions' levels of difficulty in English for Palestine 12th grade English student's textbook in terms of their categorization according to Barrett's reading comprehension higher thinking skills taxonomy. Through descriptive analysis, the researcher found that the largest proportion of the questions in the 12th grade textbook was literal level questions represented by around 60% of the textbook total number of questions exceeding the syllabus objectives with 29.9%. The reorganization, inferential, and appreciation questions were under represented compared to the syllabus objectives percentages. Only the evaluation questions were compatible with higher thinking skills Taxonomy as projected by the syllabus. Thus, the researchers recommended incorporating these findings in the student's textbook to simulate the syllabus percentages.

In brief, analyzing the ENE items is a process that sheds some light on the strengths and weaknesses of listening and reading comprehension texts and tests and their classifications of LOTS and HOTS. This study thus is aimed to answer these questions:

1. To what extent do the questions in the ENE 2013/2014 academic year include literal, reorganization, inferential, evaluation, and appreciation comprehendsion which reflect the students' LOTS and HOTS?
2. Are the proportions of items assessing students' LOTs and HOTS consistent among the twenty test packages in the ENE of 2013/2014 academic year?

Thus, the purpose of this study was to investigate the nature of questions used in the ENE for Senior High School students based on Barrett Taxonomy and its efficacy to develop the 12th grade students linguistically, mentally and intellectually. The researcher formulated a checklist of criteria for evaluating LOTS and HOTS in the ENE of 2013/2014 academic year and identified the proportions of both thinking skills levels in the listening and reading comprehension questions as well as writing performance item in the ENE, and compared the consistency of the number of items assessing students' LOTS and HOTS among the twenty test packages in the ENE. As regulated in Education National Standard Organization Regulation No. 0020/P/BSNP/I/2013, the 20 packets of the test items are professionally designed by the test designers to reflect the same table of specifications listed in Education National Standard Organization Regulation No. 0019/P/BSNP/XI/2012 that share the same level of difficulty, quality, and validity.

The roles, the importance, and the issue of authenticity of ENE were not discussed in detail as they are beyond the scope of this research. Due to the constraint of time and finance, it was not possible to investigate the issue of test validity, reliability, the level of difficulty, and the item discriminability, but only to concentrate on specific relevant questions as stated previously.

Methods

Sources of Data and Data

The sources of data were the twenty packages of the English National Exam for Senior High School students of 2013-2014 academic year. Qualitative data were taken from 1,000 test items accumulated from 20 test packages, each of which is administered to different student taking the examination. Each package contained 50 test items comprising 15 listening and 31 reading questions and 4 writing questions

with four alternatives supplied in each item.

Instruments

The major instrument in conducting this study is the researcher herself. She developed a tool called categorical content analysis to collect, describe and analyze data regarding the availability of LOTS and HOTS in the listening and reading exercises of the ENE in the light of the suggested checklist in the analysis card. To ensure the validity of the content analysis card, it was shown to some experts so that the researcher could benefit from their comments and suggestions for further modifications. Having confirmed the final version of the checklist, the writer divided the number of coded points into five categories i.e.; literal, reorganization, inferential, evaluation, and appreciation as shown in Table 2[2].

The number of coded points in this table refers to the Quick Reference Outline of Barrett Taxonomy (see Appendix 1). It is explained that the domain of Literal Comprehension consists of recognition comprehension comprising six points and recall comprehension comprising six points which add up to twelve points. The second domain, Reorganization Comprehension, consists of four points, while Inferential Comprehension domain consists of eight points. Evaluation Comprehension consists of five points and Appreciation Comprehension comprises four points.

Data Collection

To gather all information needed, the researchers collected all suitable documents that are available. The documents collected in this research were the coding sheet of comprehension questions analysis, the twenty test packages of English National Examination for Senior High School in the 2013/2014 academic year obtained from schools, and the table of specifications listed in Education National Standard Organization Regulation No. 0019/P/BSNP/XI/2012, a document which is publicly available on the Internet.

Data Analysis

To analyze the test items, the researcher used a coding sheet to classify the test items of ENE into the questions' levels of comprehension based on Barrett Taxonomy. First of all, the selected examination paper samples were sorted by assigning numbers from 1 to 20 to the papers. Then, 1,000 question items asking students' LOTS and HOTS were identified and put into several categories within the content analysis card. Sentences and concepts that make up questions in exam papers are discoverable using content analysis method.

[2] Table 2, p. 17

Trustworthiness

Building trustworthiness in this study was conducted through the help of inter-rater reliability. It is meant to assure that the result of the study is reliable and excludes any bias or the researchers' subjectivity. To achieve this, the researchers invited three raters to code the qualitative data into various categories, i.e. levels of comprehension and thinking skills. The first rater is a prolific writer and an expert in educational field and the rest are Senior High School teachers who pursue their postgraduate education at Widya Mandala Catholic University. The writer chose them due to their expertise and experience in teaching English.

A sample of test package was randomly chosen to be analyzed by the raters independent from one another. She provided the raters with the criteria prepared for evaluating the levels of comprehension questions that have been reviewed by experts. Question terms in the question base like *who, what, where, when, how, express, define, summarize, compare, plan, arrange, distinguish, show, conclude, find,* etc. have been taken into account in determining the question levels. She discussed later with them how to conduct the analysis.

Findings

Questions Requiring Students' Levels of Thinking Skills

Data for the number of questions asking students' LOTS and HOTS were obtained from all the questions in the twenty test packages of ENE. In order to show how the data were codified and analyzed, some part of the total data was chosen as an illustration. For this reason, some items of the English National Exam (ENE) of the 2013 – 2014 academic year are presented as an example.

This text is for questions 16 and 17.

Dear Big Meal's representative

I'm writing to infrom you that I had a negative bad experience at your location in Columbus, New Jersey on August 4. My receipt number is 512, and the person who handled my order was Alex.

First of all, I recognize that you, as the reader of this letter, are not responsible for my bad experience, but I am still upset about the situation.

I went to the drive through and odered seven meals with no pickles. When I received my order, I checked that all of the sandwiches and fries were in the bag paid and drove away. When I got home, I realized my number seven had pickles on it. I'm allergic to pickles, and I didn't want to waste the sandwich, so I drove back to the drive through to explain the situation and get it fixed.

I feel very disappointed with this interaction, as I usually enjoy my experiences at your restaurant. To fix this situation, I would like a coupon for a free meal of my choice. I think an apology from Alex is also appropriate.

Please contact me at 555.555.5555 or email me back at jkorkell@email.com. I would like this situation to be resolved so I can continue to be a loyal Big Meal's patron.

Best,

Jim Korkell

16. What is the letter about?
    A. Applying for a job.
    B. Complaining bad service.
    C. Ordering a certain item.
    D. Inquiring Mr Jim Korkell.
    E. Reserve for a meal.

The correct choice is B. Answering this question does not need a higher order of thinking because this question only needs locating or identifying explicit facts or detail requiring literal comprehension. Therefore, it is codified as 1.1.1 (Recognition of details) which belongs to literal comprehension.

DOKUMEN NEGARA
SANGAT RAHASIA

16

Bahasa Inggris SMA/MA IPS

The text is for numbers 40 and 41.

Beggars have become a big problem for us today. They come as street musicians, street boys, "sick" people, "lost" people, or just beggars. As their number is getting bigger, the municipal government feels the need to set a regulation to ban beggars. Many people support this.

They say that begging makes people lazy and bad survivors. They are like parasites. Criminals take advantage of their existence. Car drivers are strong-armed in crossroads, motorbikes are seized, trucks are hijacked, etc. A man in a rural area takes them to the city with his truck in the morning and pick them up in the afternoon. They have made an agreement to share what they get. Some children are reported to have been kidnapped not for ransom. They are forced to be beggars.

Some people, however, say that we must help beggars. They become beggars because they have no choice. What they get everyday is only enough for buying food. Being a beggar is better than being a thief or a robber. So it is a high time to apply their religious teaching to care for others. In addition, what they do is to help the government to check crime-rates.

Despite the controversy of their existence, beggars continue to color the life of urban people.

40. What do you think about the man who transports beggars from their villages to the city?
    A. He is generous.
    B. He is exploitative.
    C. He is a travel agent.
    D. He is their protector.
    E. He is doing business.

41. "Car drivers are strong-armed in crossroads". (Paragraph 2).

    The underlined words is closest in meaning to ....
    A. forced to give money or other valuables
    B. helped to overcome traffic problems
    C. suggested to always be careful
    D. demanded to obey traffic rules
    E. directed to the correct paths

The correct choice for number 40 is B. The question is codified as 4.5 (Judgments of Worth, Desirability and Acceptability) which is covered in evaluation comprehension. Questions of this nature

call for judgments based on the reader's moral code or his or her value system.

Furthermore, the correct choice for number 41 is A. To answer this question, the learners have to grasp the meaning by translating and interpreting. In other words, the students, in this instance, are asked to infer literal meanings from the author's figurative use of language. Thus it was codified as 3.8 (Interpreting Figurative Language) which belongs to Inferential comprehension level.

Evidently, these findings confirmed that the levels of comprehension questions in the English National Examination vary. There are five categories of comprehension levels proposed in Barrett Taxonomy, namely literal, reorganization, inferential, evaluation, and appreciation, as can be observed in the following table 3[3].

The above analysis of the ENE comprehension questions for the Senior High School level reveals that the total number of the questions (1,000 items) was distributed over the Barrett Taxonomy. It is obvious that, of the whole test packages, the items categorized into literal level represented around 68.6% of the total number of the questions. Meanwhile, the questions belonging to reorganization comprehension came to occupy a percentage of 20.8. This indicates that the questions asked in ENE were mostly in the low level of comprehension or lower order thinking skills (LOTS).

On the other hand, only few of the question items which promoted students' HOTS were available in the ENE, for instance, the inferential level only reached 10.3%. The test was also not enriched sufficiently with the evaluation comprehension since it only reached 0.3%, and the appreciation 0. This shows that there is a sign of deficiency in these three comprehension levels.[4]

As shown in Table 5[5], the distribution of each comprehension skills tested indicates the same result with the distribution of the total number of questions. In listening comprehension questions, a large part of the questions (73.3%) was seen at the level of 'Literal Comprehension'. On the other hand, the amount of questions asking the students' comprehension skills such as understanding and interpretation of the text, establishing the relationship between events containing 'Inferential Comprehension', 'Reorganization', and 'Evaluation' fields was found to be lower. Similarly, in reading comprehension questions, the majority of test items (65.5%) are also at the level of 'Literal

[3] Table 3, Pp.18-19

[4] Table 4, p. 20

[5] Table 5, p.21

Comprehension'. The writing performance item show no difference from the other two skills that literal comprehension dominates the test items (75%).

Interraters' Disagreement

For ensuring the reliability of the study, three inter raters were included in the study to examine the comprehension questions in the exam papers. Questions were addressed independently by each rater and were made the distribution of Barrett's Taxonomy Sublevels. Training on how to carry out the categorical content analysis using the data collection instruments was given to the raters in order to perform the analysis. Pertinent and relevant examples were provided. After several discussions on the procedural and conceptual issues of the instrument, a particular period of time was given to categorize the test items using the coding sheet.

The opinions of three raters included in the study were coded for each question using comparative analysis. The findings of the review (code information) were subject to an analysis of the reliability of the code. To determine inter-rater reliability, the researcher used the following formula (Miles & Huberman, 1994):

$$\text{Reliability} = \frac{\text{Number of agreements}}{\text{Total number of agreements + disagreements}} \times 100\%$$

The coding of the 50 question items resulted in approximately 80% agreement (coding agreement on 40 of 50 items in one document sample of test packages). There were disagreements and agreements with some concepts particularly on the categorization of items into the suitable domains. After initially comparing the levels, differences on 10 examples of questions were resolved by discussing the criteria contained in Appendix 6 and the rationale used by each rater to code each data source. Since there was 80% or high agreement between the coders on the 50 questions, the researcher proceeded to code the remaining questions alone. High inter-rater reliability provided increased confidence in coding consistency (Miles & Huberman, 1994).

Anatomy of ENE Test Packages

Of all the twenty packages[6], it was discovered that listening and writing sections had similar questions and options.

[6] Table 6, p. 22

The listening comprehension was required in 15 items of questions, while writing performance was covered in 4 items. It was also discovered that items which assess students' writing skill are more likely to cross over into the domain of assessing reading due to some reasons. First, the words and phrases which serve as the options of the stem are presented in the form of a multiple-choice test. The students are not required to write down answers which enable teachers to assess their correct spelling or the students' ability to organize and develop ideas logically. Second, the indicators of students writing skills mentioned in the table of specifications merely cover the students' competence to arrange jumbled sentences into a paragraph and to fill in the blanks of cloze test. According to Brown (2004, pp. 201-210) these types of assessment tasks are classified into assessing interactive reading; cloze test and sentence-ordering task.

The reading section varied from one test package to others. However, the reading passages in each of those twenty test packages were not completely different since the researcher found out that there were 3 sets of test packages containing almost 80% similar reading passages (11 texts out of 13). Thus, the researcher classified the twenty test packages into 7 groups of test packages since the rest also adopted the pattern of 3 sets of test packages in which most reading passages were similar. The classification is illustrated in the table below.

Furthermore, in the ENE, none of the items have asked students' appreciation level of comprehension. This reveals that the exercises need more varied questions that enable students to elicit emotional responses to content, plot or theme, sensitivity to various literary genres, identification with characters and incidents, reaction to author's use of language, and response to generated images. It is the top skill of Barrett Taxonomy.

Discussions

Based on the findings, the majority of the questions focused primarily on the comprehension level of literal and reorganization (LOTS) than HOTS. LOTS items comprised 87.4% and HOTS 10.6%. It reveals that students' HOTS were not well-treated or rather neglected. It is worth noting that the lack of these items categorized into inferential, evaluation and appreciation means the negligence to include the students' higher order thinking skills. Concerning these findings, it can be said that the comprehension questions in all of the ENE test packages needed to be enriched with more HOTS such as the inferential, evaluation and appreciation comprehension levels which had the least

share in the ENE items if compared with the other two levels of comprehension (literal and reorganization). In other words, more evaluative questions should be provided so that students would have the opportunity to express their opinions, feelings, and attitudes which pave their way to be creative and innovative thinkers.

The negative impact of the test design which does not stimulate learners to optimize their critical thinking is a serious concern. Bachman & Palmer (1996, p. 18) define impact in terms of the various ways a test's use affects the society, an educational system, and the individuals within them. The consequences of the test design are extremely serious and are burdened not only to students, but also to teachers. Students do excessive amount of drilling for test practices. Consequently, students experience psychological distress. They feel worried and anxious of failing to pass the test. Besides, after taking the test, which fits the description of the high-stakes testing, students do not feel satisfied since their full potential are not well explored. Moreover, teachers have been discouraged to teach in engaging and meaningful ways. They are forced to sacrifice their creative, innovative, meaningful, and engaging lessons to allow time for students to practice the test drills, which mostly focus on the Lower Order of Thinking. Lessons are adjusted towards memorizing the information needed to answer the multiple-choice paper-and pencil exams.

Meanwhile, there is a lack of progression from the lower cognitive skills to the higher ones. Ideally, the question items must be arranged in a linear fashion. The items which contain literal comprehension must come first and gradually followed by comprehension questions asking students' higher level of thinking. However, in the anatomy of ENE, the writer found out that this principle of language testing is ignored. The test packages analyzed in this study were made for Senior High School students majoring in science. When the writer compared them to those for students majoring in social studies, she found out that almost all the questions are similar but the order of questions in each test package was different. They are not arranged in a systematic order from the simplest to questions that require the most complicated answers.

On the other hand, all these test packages are evenly distributed throughout Indonesia, leaving no difference in remote area or big cities. For instance, in East Java, students in Sampang receive the same tests as those in Surabaya. It creates a big gap of students' achievement because the actual capability of schools in rural areas to meet the demands of national

exam vary greatly from those in urban areas.

In relation to criteria of measurement qualities of test suggested by Bachman & Palmer (1996, p. 18) which describes a good language test usefulness, the ENE items demonstrate some criteria such as construct validity, authenticity, and practicality. The questions used in the test are relevant and representative of the skills measured in the table of specifications used for 2013/2014 academic year which refers to that listed in Education National Standard Organization Regulation No. 0019/P/BSNP/XI/2012.

The test also shows its authenticity through the use of the target language. The listening materials are spoken by native speaker and the reading texts demonstrate to students the real-world context of the language use such as advertisement, movie review, book review, various types of letters, and articles.

In terms of practicality, which can be observed from several aspects: (1) economy of time, money, and labor; (2) ease of administration and scoring; and (3) ease of interpretation (Nation & Newton, 2009, p. 166), the ENE design demonstrates all the aspects. It is administered in a multiple choice format since it is an efficient and effective way to assess a wide range of skills. It is also easier to score due to objective assessment. In fact, if done well, multiple choice format can measure whether students "understand at the most explicit literal level, make pragmatic inferences, understand implicit meanings and summarize or synthesize extensive sections of tests".

The overall findings of this study demonstrated that higher order cognitive skills in ENE items are not well covered, not well treated nor well distributed. To illustrate, out of the 1,000 questions analyzed, only 106 items ask students' higher order thinking skills. This is ironic since at their age, students of Senior High School are demanded to be able to cope with the development of technology as well as the creative industry. Consequently, students need to sharpen their knowledge and insight, exercise their minds to think critically, and learn to communicate effectively so that they can survive to deal with the challenges of the 21st century and the era of Asian Economic Community (AEC). It is in line with Trilling & Fadel who point out that there will be a rising demand of workers who can fill in the jobs that involve higher levels of knowledge and applied skills like "expert thinking and complex communicating" (2009, p. 8).

In consequence, raising the awareness among teachers and educators as well as the society that curriculum and educational

processes are responsible for building learner's critical thinking is deemed very crucial. If the ENE is designed to test students' HOTS, most teachers' and students' activities in the classroom will be oriented toward improving these skills. In turn, this practice will be beneficial for students for their whole academic lives. Otherwise, if the test are dominated with questions asking the students' LOTS, students will be low achievers who are merely capable of focusing on lower order thinking skills (LOTS). This is in line with Jacob in Sukyadi & Mardiani (2011) who states that high school national graduation exams increased the rate of drop outs and hinder the development of higher order thinking skills.

As a result, an effort from the test designers should be exerted to provide items that cover the missing parts of the test related to these three comprehension levels. Otherwise, the question items do not satisfy competent students who need challenging questions to promote their thinking abilities because they primarily focus on the lower skills such as literal and reorganization. In other words, more emphasis should be given to the questions asking students' higher order thinking skills.

Conclusions and Suggestions

The results of this study indicate the presence of almost all levels of thinking in English National Exam (ENE) items in Indonesia, except "Appreciation" which is the highest level of thinking in Barrett taxonomy. It is obvious that literal which is included in Lower Order Thinking Skills, among all levels of comprehension has the highest percentage; its percentage equals 68.60%. Accordingly, the order of thinking levels for ENE items from the one with the highest percentage, toward the lowest one is as follows: Literal (68.60%), Reorganization (20.80%), Inferential (10.3%), Evaluation (0.3%), and Appreciation (0%). In other words, the majority of the questions focused primarily on the comprehension level of literal and reorganization (LOTS) than inferential, evaluation, and appreciation (HOTS) as LOTS items comprised of 87.4% and HOTS 10.6%.

Accordingly, based on the results of this study, it can be concluded that Lower Order Thinking Skills (LOTS) are the main concern of ENE items. This finding reveals that there is still much room for ENE to be the driving force in the effort to make learners critical thinkers. It must be accompanied by classroom exercises in all English skills which require students' HOTS. Furthermore, it is clear that those crucial principles necessary for

constructing good test items are not met in ENE items in Indonesia.

Recommendations for future practice and research include the following:

1. It is recommended that the test designers should modify the question items in ENE to include higher order thinking skills.

2. The Ministry of Education instructs the test developers to coordinate with curriculum developers to create alignment between the ENE comprehension questions with the curriculum to ensure the reduction of literal level questions and increase the questions requiring comprehension levels which belong to HOTS.

3. English supervisors are recommended to prepare enrichment materials that provide teachers with more exercises that cover higher order thinking skills. In addition, they should hold more workshops to train the English teachers how to develop and enhance students' thinking skills.

4. Other researchers need to conduct studies related to the current one in other NE items to see to what extent the higher levels of thinking were more reflected.

To ensure students success and prepare them to face the challenges in 21st century, it is very crucial to train them to have creative and critical thinking. One of the ways to reach the purpose is by providing them intensive exercises to answer questions requiring their higher order thinking skills such as those belong to inferential comprehension, evaluation, and appreciation level. In this case, assessment and evaluation practices of teachers are of great importance.

© Ummu Lathifah Ahmad 2016

Ummu Lathifah Ahmad received a bachelor's degree in English Education from Surabaya State University (2009) and a master's degree in the same field from Widya Mandala Catholic University, Surabaya, Indonesia, in (2015). Since 2011, she has joined the East Java Provincial Education Department, where she is working as a staff in Budget Planning Division. She is currently a Lecturer at STID Al Hadid Surabaya. Her current research interests include educational policies, English language teaching, early childhood care and education.

Suggested reference format for this article:

Ahmad, U. L. (2016, November). Senior High School English National Examination and Thinking Skills. *Beyond Words, 4*(2), 168 - 190. Retrieved from journal.wima.ac.id/indexed.php/BW

References

Armbruster, B., & Ostertag, J. (1993). *Learning from Textbooks : Theory and Practice.* (B. K. Britton, A. Woodward, & M. Brinkley, Eds.) New Jersey: Lawrence Erlbaum Associates Inc.

Bachman, L. F., & Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Brown, H. D. (2004). *Language Assessment (Principals and Classroom Practices).* Pearson: Longman.

Brown, J., & Rodgers, T. (2002). *Doing Second Language Research.* Oxford: Oxford University Press.

Humos, O. A. (2012). An Evaluative Analysis of Comprehension Questions' Level of Difficulty: A case of 12th Grade Palestinian English Student's Textbook. *An - Najah Univ. J. Res. (Humanities), 26*(3), 767-788.

McDavitt, D. (1993). *Teaching for Understanding: Attaining Higher Order Learning and Increasing Achievement through Experimental Instruction.* Unpublished Thesis.

Miles, M. B., & Huberman, M. A. (1994). *Qualitative Data Analysis : An Expanded Sourcebook.* California: Sage Publications.

Nation, I., & Newton, J. (2009). *Teaching ESL/EFL Listening and Speaking.* New York: Routledge.

Razmjoo, S. A., & Madani, H. (2013, November). A Content Analysis of the English Section of University Entrance Exams Based on Bloom's Revised Taxonomy. *International Journal of Language Learning and Applied Linguistics World, 4*(3), 105-129.

Seif, A. A.-Q. (2012). *Evaluating the Higher Order Thinking Skills in Reading Exercises of English for Palestine Grade 8.* The Islamic University-Gaza, Department of Curricula and Methodology. Gaza: The Islamic University-Gaza.

Sukyadi, D., & Mardiani, R. (2011, June). The Washback Effect of the English National Examination (ENE) on English Teacher's Classroom Teaching and Students' Learning. (A. H. Nugroho, E. Kuntjara, J. M. Djundjung, & P. F. Handojo, Eds.) *Kata, 13*(1), 96-111.

Tomei, L. (2005). *Taxonomy for the Technology Domain.* London.

Trilling, B., & Fadel, C. (2009). *21st Century Skills: Learning for Life in Our Times.* San Fransisco: Jossey-Bass.

Wilson, V. (2000). Can thinking skills be taught?

Tables

Table 1
Bloom's, Barrett Taxonomy and Two-Level Thinking Skills Model

Bloom et al. (1956)	Barrett (1979)	Two-Level Thinking Skills Model: LOTS and HOTS
Knowledge	Literal recognition or recall	Lower-Order Thinking Skills
Comprehension	Reorganization	
Application	Inference	Higher-Order Thinking Skills
Analysis		
Synthesis		
Evaluation	Evaluation	
	Appreciation	

Table 2
The number of coded points in each domain

No.	Domains	Number of coded points
1.	Literal comprehension	12
2.	Reorganization comprehension	4
3.	Inferential comprehension	8
4.	Evaluation comprehension	5
5.	Appreciation comprehension	4
	Total	33

Table 3

Recapitulation of Question Types Based on Barrett Taxonomy

Item Number	Test Package Number																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
2.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
3.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
4.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
5.	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
6.	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
7.	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
8.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
9.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
10.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
11.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
12.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
13.	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
14.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
15.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
16.	R	R	R	R	R	R	R	R	R	E	E	E	R	R	R	R	R	R	R	R
17.	L	L	L	L	L	L	L	L	L	I	I	I	L	L	L	L	L	L	L	L
18.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
19.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
20.	L	R	R	R	R	R	R	R	R	L	L	L	L	L	L	I	I	I	R	R
21.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	I	I
22.	L	L	L	L	L	L	R	R	R	L	L	L	L	L	L	L	L	L	L	L
23.	R	R	R	L	L	L	R	R	R	L	L	L	R	R	R	R	R	R	R	R
24.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L	L

Item Number	Test Package Number																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
25.	L	L	L	L	L	L	L	L	L	R	R	R	L	L	L	L	L	L	L	L
26.	I	I	I	I	I	I	R	R	R	R	R	R	R	R	R	R	R	R	R	R
27.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
28.	L	L	L	L	L	L	L	L	L	R	R	R	L	L	L	L	L	L	L	L
29.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
30.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
31.	I	I	I	L	L	L	I	I	I	L	L	L	R	R	R	L	L	L	L	L
32.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
33.	L	L	L	L	L	R	L	R	R	R	R	I	I	I	R	L	L	R	R	R
34.	I	I	R	R	R	L	I	L	L	L	L	R	R	R	L	L	L	L	L	L
35.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
36.	R	L	I	L	R	R	L	L	L	L	R	L	R	R	R	R	L	R	R	R
37.	L	R	R	R	L	L	L	R	L	R	L	R	L	I	L	L	L	L	I	L
38.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
39.	L	L	L	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
40.	L	L	L	L	L	L	I	I	I	L	L	L	L	L	L	L	L	L	R	R
41.	R	R	R	L	L	L	R	R	R	L	L	L	L	L	L	L	L	L	L	L
42.	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
43.	L	R	L	R	R	R	R	R	L	R	R	R	R	R	R	R	R	R	R	R
44.	L	L	L	L	R	L	L	L	L	L	R	R	L	L	L	L	L	L	L	L
45.	I	R	I	R	R	R	R	R	I	L	R	R	L	L	L	R	L	L	L	L
46.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
47.	I	I	I	I	I	I	I	I	I	R	R	R	R	R	R	R	R	R	R	R
48.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
49.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L
50.	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L	L

Table 4.

Comparison between Each Test Package Based on Barrett Taxonomy

TEST PACKAGE	Literal		Reorganization		Inferential		Evaluation		Appreciation	
	Number of items	%	Number of items	%	Number of items	%	Number of items	%	Number of items	%
1	36	72%	6	12%	8	16%	0	0	0	0
2	35	70%	8	16%	7	14%	0	0	0	0
3	36	72%	7	14%	7	14%	0	0	0	0
4	36	72%	9	18%	5	10%	0	0	0	0
5	35	70%	10	20%	5	10%	0	0	0	0
6	36	72%	9	18%	5	10%	0	0	0	0
7	32	64%	11	22%	7	14%	0	0	0	0
8	31	62%	13	26%	6	12%	0	0	0	0
9	33	66%	10	20%	7	14%	0	0	0	0
10	35	70%	10	20%	4	8%	1	2%	0	0
11	33	66%	12	24%	4	8%	1	2%	0	0
12	32	64%	12	24%	5	10%	1	2%	0	0
13	35	70%	11	22%	4	8%	0	0	0	0
14	34	68%	11	22%	5	10%	0	0	0	0
15	36	72%	11	22%	3	6%	0	0	0	0
16	34	68%	12	24%	4	8%	0	0	0	0
17	36	72%	10	20%	4	8%	0	0	0	0
18	34	68%	12	24%	4	8%	0	0	0	0
19	33	66%	12	24%	5	10%	0	0	0	0
20	34	68%	12	24%	4	8%	0	0	0	0
TOTAL	**686**	**68.6%**	**208**	**20.8%**	**103**	**10.3%**	**3**	**0.3%**	**0**	**0**

Table 5

The Distribution of Comprehension Questions

Domain	The Distribution of Total Questions		The Distribution of Listening Questions		The Distribution of Reading Questions		The Distribution of Writing Performance	
	f	%	*f*	%	*f*	%	*f*	%
Literal	686	68.5	220	73.3	406	65.5	60	75
Reorganization	208	20.8	20	6.7	168	27.1	20	25
Inferential	103	10.3	60	20	43	6.9	0	0
Evaluation	3	0.3	0	0	3	0.5	0	0
Appreciation	0	0	0	0	0	0	0	0
Total	1,000	100	300	100	620	100	80	100

Table 6

The Classification of Test Packages Based on the Similarity in Test Items

Sections	Groups of Test Packages						
	1 (TP 1-3)	2 (TP 4-6)	3 (TP 7-9)	4 (TP 10-12)	5 (TP 13-15)	6 (TP 16-18)	7 (TP 19-20)
Listening	Exactly the same question items (Item no. 1 – 15)						
Reading	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47) **TP 1:** 33 – 35 (Christiano Ronaldo) 36 – 38 (Bali) 43 – 46 (Biodiesel) **TP 2:** 33 – 35 (Christiano Ronaldo) 36 – 38 (Flashmob) 43 – 46 (New glass bottle) **TP 3:** 33 – 35 (Dr. Abdurrachman Saleh) 36 – 38 (Flashmob) 43 – 46 (Biodiesel)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47) **TP 4:** 33 – 35 (Dr. Abd. Saleh) 36 – 38 (Thailand) 43 – 46 (New glass bottles) **TP 5:** 33 – 35 (Dr. Abd. Saleh) 36 – 38 (Bali) 43 – 46 (Memory) **TP 6:** 33 – 35 (Venus Williams) 36 – 38 (Bali) 43 – 46 (New glass bottles)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47) **TP 7:** 33 – 35 (Christiano Ronaldo) 36 – 38 (Tangerang) 43 – 46 (Food) **TP 8:** 33 – 35 (Messi) 36 – 38 (Flashmob) 43 – 46 (Food) **TP 9:** 33 – 35 (Messi) 36 – 38 (Tangerang) 43 – 46 (Biodiesel)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47) **TP 10:** 33 – 35 (Venus Williams) 36 – 38 (Thailand) 43 – 46 (Fossil fuels) **TP 11:** 33 – 35 (Venus Williams) 36 – 38 (Norway) 43 – 46 (Memory) **TP 12:** 33 – 35 (Louis Lionel Messi) 36 – 38 (Thailand) 43 – 46 (Memory)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47) **TP 13:** 33 – 35 (Louis Lionel Messi) 36 – 38 (Norway) 43 – 46 (Solar energy) **TP 14:** 33 – 35 (Louis Lionel Messi) 36 – 38 (Damaged roads) 43 – 46 (Fossil fuels) **TP 15:** 33 – 35 (Neymar da Silva) 36 – 38 (Norway) 43 – 46 (Fossil fuels)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47) **TP 16:** 33 – 35 (Kaka) 36 – 38 (Ragunan Zoo) 43 – 46 (Food) **TP 17:** 33 – 35 (Kaka) 36 – 38 (Tangerang) 43 – 46 (Butterflies) **TP 18:** 33 – 35 (Messi) 36 – 38 (Ragunan zoo) 43 – 46 (Butterflies)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47) **TP 19:** 33 – 35 (Neymar da Silva) 36 – 38 (Damaged roads) 43 – 46 (Butterflies) **TP 20:** 33 – 35 (Neymar da Silva) 36 – 38 (Ragunan zoo) 43 – 46 (Solar energy)
Writing	42 (The Crocodile and Monkey)	42 (Red Riding Hood)	42 (The Smartest Parrot)	42 (The Elephant)	42 (Batara Gum Sahala)	42 (Cleopatra)	42 (The Frogs)
	Exactly the same question items (Item no. 48 –50)						