Chapter 4

Results and Discussion

This chapter answered the questions of the study by presenting the findings of the used tool (the categorical content analysis using coding scheme) and the anatomy of the test packages.

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 2013 – 2014 academic year. 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.

The text is for question 39.

The depletion of fossil fuels such as coil, oil and gas leads to seek for some alternate power sources, and nuclear power is the only power source that can meet the electricity demands in the future. However, nuclear power has some advantages and disadvantages. Of the advantages, nuclear is efficient. One kg of fuel of uranium gives energy equivalent 3,000 tonnes of high grade coal. Next, nuclear power plants require a little space compared to thermal power plant for the same MW output. Also, nuclear fuel is available in plenty amount all over the world. Therefore, fuel supply to plants will be continuing for hundreds of years. But, some disadvantages of nuclear plants are also obvious. In nuclear plants safety is primary concern rather producing electricity. There is significant risk of leakage of radiation in case of any accident. The fission by products released are generally radio-active and pollute the land, water, atmosphere and other natural resources. It requires large water mass for cooling purposes. Therefore, the plant should be near to a sea or river with the risk of contaminated water. Finally, it requires a large area around the plant to be isolated from living, almost 5 km in radius.

39. Nuclear radiation can be a threat to the environment by way of ...
    A. leaking it in an accident.
    B. emitting greeenhouse gasses.
    C. using water to cool in nuclear reaction.
    D. needing more uranium from the nature.
    E. requiring areas large enough to build the plant.

The correct choice is A. Since the students are asked to condense the selection using direct or paraphrased statements from the selection, the question is codified as 2.3.

(summarizing) and categorized into reorganization.

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.

Responding to the six research questions, the following findings, derived from scrutinizing words and phrases presented in the questions, were developed in the coding scheme. 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.

Table 4.1.
Recapitulation of Comprehension Question Types Based on Barrett Taxonomy

Item Number	**Test Package Number**																			
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**	**XI**	**XII**	**XIII**	**XIV**	**XV**	**XVI**	**XVII**	**XVIII**	**XIX**	**XX**
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

Item Number	Test Package Number																			
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
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

Item Number	Test Package Number																			
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
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
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

Item Number	Test Package Number																			
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
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

Item Number	Test Package Number																			
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
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

Notes:

L = Literal, R = Reorganization, I = Inferential, E = Evaluation, A = Appreciation

Table 4.2. Comparison between Each Test Package Based on the Number and Percentage of Comprehension Question Types 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	%
I	36	72%	6	12%	8	16%	0	0	0	0
II	35	70%	8	16%	7	14%	0	0	0	0
III	36	72%	7	14%	7	14%	0	0	0	0
IV	36	72%	9	18%	5	10%	0	0	0	0

TEST PACKAGE	Literal		Reorganization		Inferential		Evaluation		Appreciation	
	Number of items	%	Number of items	%	Number of items	%	Number of items	%	Number of items	%
V	35	70%	10	20%	5	10%	0	0	0	0
VI	36	72%	9	18%	5	10%	0	0	0	0
VII	32	64%	11	22%	7	14%	0	0	0	0
VIII	31	62%	13	26%	6	12%	0	0	0	0
IX	33	66%	10	20%	7	14%	0	0	0	0
X	35	70%	10	20%	4	8%	1	2%	0	0

TEST PACKAGE	Literal		Reorganization		Inferential		Evaluation		Appreciation	
	Number of items	%	Number of items	%	Number of items	%	Number of items	%	Number of items	%
XI	33	66%	12	24%	4	8%	1	2%	0	0
XII	32	64%	12	24%	5	10%	1	2%	0	0
XIII	35	70%	11	22%	4	8%	0	0	0	0
XIV	34	68%	11	22%	5	10%	0	0	0	0
XV	36	72%	11	22%	3	6%	0	0	0	0
XVI	34	68%	12	24%	4	8%	0	0	0	0

TEST PACKAGE	Literal		Reorganization		Inferential		Evaluation		Appreciation	
	Number of items	%	Number of items	%	Number of items	%	Number of items	%	Number of items	%
XVII	36	72%	10	20%	4	8%	0	0	0	0
XVIII	34	68%	12	24%	4	8%	0	0	0	0
XIX	33	66%	12	24%	5	10%	0	0	0	0
XX	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**

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).

In addition, few of the question items which promoted students' HOTS were available in the ENE, specifically, the inferential level which only reached 10.3%. These items improve students' ability in reading critically, guessing meaning through context. However, the writer argues that there is still an urgent need to focus more on the other two skills; evaluation and appreciation. The test was 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 two comprehension levels.

Table 4.3.
The Distribution of Listening Comprehension, Reading Comprehension, Writing Performance and Total Questions According to Barrett's Taxonomy

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

As shown in Table 4.3, 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, percentage of exploring students understanding and 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 Comprehension’. The writing performance item show no difference from the other two skills that literal comprehension dominates the test items (75%).

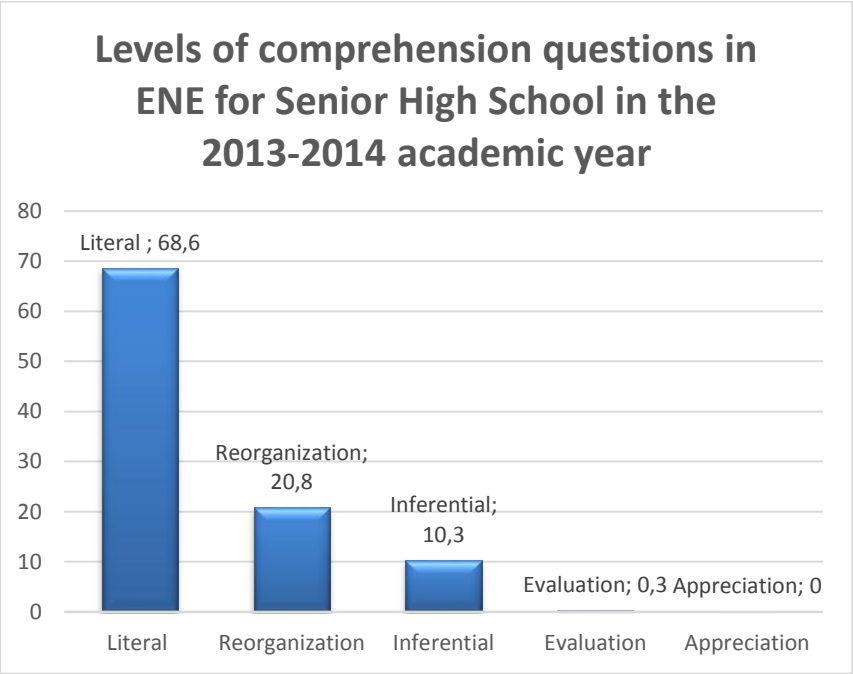

Figure 4.1. Total Average of ENE Items for SHS in the 2013 – 2014 Academic Year

Generally speaking, the majority of the questions focused primarily on the comprehension level of literal and reorganization (LOTS) than HOTS as LOTS items comprised 87.4% and HOTS 10.6%. It reveals that students' HOTS were not well-treated or rather neglected. 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).

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. Consequently, 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.

Interraters' Disagreement

For the reliability of the study, three inter raters were included in the study for the examination of comprehension questions in the exam papers. Questions 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 questions resulted in approximately 80% agreement (coding agreement on 40 of 50 examples of questions 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 5 and the

rationale used by each rater to code each data source. Since far greater than 80% reliability was achieved, it was not necessary to recode or provide additional items for scoring. High inter-rater reliability provided increased confidence in coding consistency (Miles & Huberman, 1994). Since there was more than 80% or higher agreement between the coders on the 50 questions, the researcher proceeded to code the remaining questions alone.

Anatomy of ENE Test Packages

In this study, a total of 20 test packages of English National Exam for science major of Senior High School students were selected. From these examination papers 1,000 questions were analyzed. Of all the twenty packages, it was discovered that listening and writing sections had similar questions and options. 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.

Table 4.4. 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)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47)	Exactly the same reading passage and question items (Item no. 16–32, 39-41, 47)	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)	TP 4: 33 – 35 (Dr. Abd. Saleh) 36 – 38 (Thailand) 43 – 46 (New glass bottles) TP 5: 33 – 35 (Dr. Abd. Saleh) 36 – 38	TP 7: 33 – 35 (Christiano Ronaldo) 36 – 38 (Tangerang) 43 – 46 (Food) TP 8: 33 – 35 (Messi) 36 – 38	TP 10: 33 – 35 (Venus Williams) 36 – 38 (Thailand) 43 – 46 (Fossil fuels) TP 11: 33 – 35 (Venus Williams)	TP 13: 33 – 35 (Louis Lionel Messi) 36 – 38 (Norway) 43 – 46 (Solar energy) TP 14: 33 – 35 (Louis	TP 16: 33 – 35 (Kaka) 36 – 38 (Ragunan Zoo) 43 – 46 (Food) TP 17: 33 – 35 (Kaka) 36 – 38	TP 19: 33 – 35 (Neymar da Silva) 36 – 38 (Damaged roads) 43 – 46 (Butterflies) TP 20: 33 – 35 (Neymar da Silva)

	36 – 38 (Flashmob) 43 – 46 (New glass bottle) TP 3: 33 – 35 (Dr. Abdurrachman Saleh) 36 – 38 (Flashmob) 43 – 46 (Biodiesel)	(Bali) 43 – 46 (Memory) TP 6: 33 – 35 (Venus Williams) 36 – 38 (Bali) 43 – 46 (New glass bottles)	(Flashmob) 43 – 46 (Food) TP 9: 33 – 35 (Messi) 36 – 38 (Tangerang) 43 – 46 (Biodiesel)	36 – 38 (Norway) 43 – 46 (Memory) TP 12: 33 – 35 (Louis Lionel Messi) 36 – 38 (Thailand) 43 – 46	Lionel Messi) 36 – 38 (Damaged roads) 43 – 46 (Fossil fuels) TP 15: 33 – 35 (Neymar da Silva) 36 – 38 (Norway)	(Tangerang) 43 – 46 (Butterflies) TP 18: 33 – 35 (Messi) 36 – 38 (Ragunan zoo) 43 – 46 (Butterflies)	36 – 38 (Ragunan zoo) 43 – 46 (Solar energy)

				(Memory)	43 – 46 (Fossil fuels)		
Wri-ting	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)						

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.

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.

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".

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.

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). 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.