A.A. Gede Yudha Paramartha (2017). The Analysis Of Multiple-Choice Test Quality For Reading III Class In English Education Department, Universitas Pendidikan Ganesha Bali, Indonesia. *Journal of Education Research and Evaluation.* Vol.1 (1) pp. 46-56.

The Analysis Of Multiple-Choice Test Quality For Reading III Class In English Education Department, Universitas Pendidikan Ganesha Bali, Indonesia

A.A. Gede Yudha Paramartha [1, *]

[1] *English Departement. Universitas Pendidikan Ganesha, Indonesia*

Abstract

This study aimed at investigating the quality of 40-item multiple-choice taken from Reading III class in English Education Department. It was carried out in order to get clearer picture and provide feedbacks on how the Critical Reading test for the new curriculum should be developed and what factors should be taken into account when developing the test. The data were taken from 24 fourth semester students in English Education Department, Unidksha, Bali who took the reading test. The data were analyzed by conducting Classical Test Theory analysis with the assistance of jMetrik software. In general, the result shows that the test was consistent but it was easy and only 40% of total items are eligible to be used. Along the findings and discussion, some feedbacks are provided for future development for the reading test.

Keywords:
Classical test theory, English, Multiple-choice test, Reading

Introduction

As a response to the release of Indonesian Qualification Framework, English Education Department, Ganesha University of Education in Bali, Indonesia has developed a new curriculum to meet the outcome standard expected by KKNI. This curriculum has been developed by English Education Department since 2015. One of the subjects in the curriculum is Critical Reading subject which aims at developing students' critical reading ability on English learning and instruction though reading various authentic reading texts. Critical Reading subject emphasizes on higher-order of thinking skill to process the information in the reading activities. Students who have higher-order of thinking skill read beyond the basic level of comprehension; they are expected to be able to analyze, synthesis, and evaluate the text at complex and deep levels (Tankersley, 2003). They are able to make critical interpretation, draw deep conclusion, relate what they have read to their background knowledge in relevant and new situations.

The most widely used thinking framework is developed by Benjamin Bloom called Bloom's Taxonomy. This taxonomy lists six major hierarchical cognitive categories from the lowest thinking skill to the highest: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluating (Anderson, 2001). Anderson then revised the taxonomy hierarchy to Remember, Understand, Apply, Analyze, Evaluate, and Create. Those six hierarchical cognitive categories can be divided into two major thinking skills: lower-order of thinking skills (LOTS: Remember, Understand, Apply) and higher-order of thinking skills (HOTS: Analyze, Evaluate, and Create).

Related to Critical Reading subject in English Education Department, the curriculum should consider the thinking skill of Analyze, Evaluate, and Create to foster students' higher-order of thinking skill. The consideration is not only for the planning and teaching and learning activity in the classroom, but also for the assessment. In reality, the reading tests used previously in the old curriculum still mostly concerned on the lower-order of thinking skill and only touched the slightest bit of higher-order of thinking skill. It happened even for the highest level of Reading subject (Advanced Reading subject) in English Education Department. Thus, it is crucial for the Critical Reading subject assessment in the newly-developed curriculum to be made with this consideration extensively.

In regard to that matter, the previous reading test needs to examined to be able to give a clearer picture about the quality and the level of thinking order assessed by the test. By doing this, it is hoped that some feedbacks and revisions can be formulated so that the quality of the reading test for the new curriculum will get better.A 40-item multiple-choice test from the previous curriculum was taken to be

* Corresponding author.
E-mail Address: yudha.paramartha@undiksha.ac.id (A.A. Gede Yudha Paramartha)

assessed. For that matter, the aim of this paper is to investigate the quality of the multiple-choice test by usingClassical Test Theory in order to provide feedbacks for future test development in the new curriculum.

Literature Reviews

Critical reading and higher-order of Thinking skill

To non-critical reader, the text that they read only offer the surface truth and information about what they read. In comparison, critical reader will not be satisfied with only the surface information; they are able to get the meaning beyond understanding (Yu, 2015). When reading a text, they understand and analyze it at the same time. Thus, the process is not only to recognize and understand the information in the text, but also it is more about analytic activity. That is why critical readers do a lot of thinking and analyzing. When someone reads critically, he combines his prior knowledge and values together with the reading materials to construct a deep interpretation about what has been read (Wallace, 2003). This idea is in-line with Yu's idea which says that non-critical reader is satisfied only with recognizing what the text says, while critical reader is more likely to go beyond what is stated in the text.

To read critically, one must actively question himself in order to recognize and analyze the information. The questions to ask while reading involves:

1. Why the author writes this?
2. What is the important information in the text?
3. Which part of the text should I analyze and summarize?
4. Do I accept the authors' arguments, opinions, or conclusion?

To answer those questions, one needs deep and thorough reading process because the meaning of the text cannot be grasp by only using the surface information of the text. When one can get deep and thorough meaning from a text, he is able to process information in the text at higher levels of thinking process (Tankersley, 2003). As stated by Anderson when he proposed the revision of Bloom's Taxonomy (Anderson, 2001), higher-order of thinking skills involve three cognitive processes, they are Analyze, Evaluate, and Create, while the lower-order of thinking skills involve remember, understand, and apply. The following is the representation of the taxonomy.

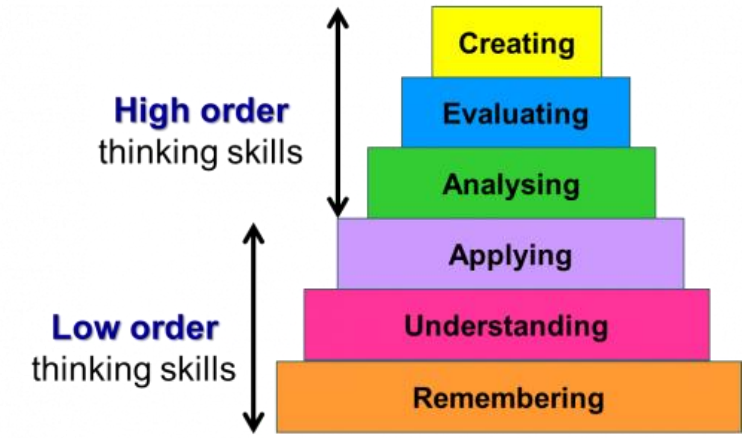

Figure 1. Bloom's Taxonomy and Thinking Skills Level

By consulting to Bloom's Taxonomy, it is clear that critical readers need cognitive process above understanding. Critical readers are able to interpret, analyze, and evaluate different aspects of the text which include context, structure, and purpose of the text (Afflerbach, 2015). Furthermore, they are able to make complex inferences by using information from the text and prior knowledge of the readers. Interpreting means constructing explicit and implicit meaning from the text in a logical and analytical way. Analyzing means making comparison with the information in the background knowledge to be able to form judgments and opinions. Evaluating means being able to distinguish important information from simply interesting information, and make judgments about ideas being read. Critical readers should use those cognitive processes while reading a text.

The learning indicators of Critical Reading syllabus which has been developed by English Education Department, Undiksha are in-line with the idea by Afflerbach (2015) about what readers should have to be called critical readers. In summary, the indicators of the Critical Reading syllabus are as follows:

1. Able to interpret the idea and meaning from a text;
2. Able to explain an argument to support and against an issue about a topic in a text;
3. Able to analyze language structure of a text;
4. Able to analyze the context of a text;
5. Able to analyze facts and opinion from a text;
6. Able to evaluate a text to see the strengths, weaknesses, and the benefit from a text.

These standardswere used as a base to assess the students' critical reading achievement. Liu (2014) reported that the majority of assessments to assess critical thinking skill (in which critical reading is one of the requirement) exclusively use selected-response items such as multiple-choice or Likert-type items. The reasons why selected-response is mainly used because it needs relatively shorter time to do and evaluate, and it is more likely to have higher level of reliability compared to constructed-response. For this project, a multiple-choice test will be developed to assess Higher-Order of Thinking Skill-based Critical Reading subject.

In recent years, researchers have tried to develop and use multiple-choice tests to assess critical reading achievement and also to assess student's higher-order of thinking skill. The followings are three research reports concerning this matter.

The first is research report by Liu (2014) about the quality of assessment format to assess critical thinking in higher education. He discovered that the majority of assessments for critical thinking are in selected-response formats such as multiple-choice or Likert-type items, and only a small number of constructed-response items are used. Moreover, Liu reported that the use of selected-response items has a problem about the inadequacy of deep information from the result of the tests, but still relevant to assess critical thinking.

Donneley (2014) explored the impact of multiple-choice test for university curriculum. This research was based on the criticism that the use of multiple-choice test can only assess surface learning achievement of the students. Her research found that the use of multiple-choice is still relevant for university students. However, she suggested that it should be blended in form of case-based. Cased-based multiple-choice test contributes to higher levels of learning and deeper information processing. Moreover, this assessment method is very useful for large class size.

A research by Hassan (2016) examined the quality of one best answer multiple-choice test for undergraduate medical education. Their research found that multiple-choice test is considered an effective tool to assess higher-order of thinking skills since it more likely has higher reliability and validity compared to other assessment formats. However, the use of multiple-choice considerably has high impact of error in assessment. Thus, they suggested that a band score should be utilized to make logical decision about students' achievement by calculating the standard error of measurement (SEM).

From the three research reports, it can be concluded that: (1) multiple-choice test is relevant to be used for university curriculum, critical thinking, and higher-order of thinking skills; (2) the use of multiple-choice should be in form of case-based format; and (3) band score should be implemented to make logical decision about students' achievement.

Classical Test Theory for Item Analysis

When developing items in an instrument, the quality of the items should be considered. It is to make sure that the test result is trustworthy and we will be confident in saying that the test reflects the real condition (or performance) of the test takers. In that regard, item analysis is very crucial to be conducted. The trustworthiness of a test will be met if a test has an intended degree of reliability and validity (Crocker & Algina, 1986). The construction of a test, the quality can be assessed by a process known as item analysis.

In classical test theory (CTT), the focus is to examine the quality of measurement by quantifying various characteristic of test items. CTT provides information about facility value, item discrimination, distractors quality,and reliability (Meyer, 2014). *Facility value* is a proportion of the students who answer the item correctly to the total number of students taking the test.The facility value ranges from 0 to 1. The larger the facility value, the easier the item is to be understood by the test taker. The following is the standard of facility value by Hopkins and Antes (1989).

Table 1. Classification for Index of Difficulty

Index	Difficulty
≥0.86	Very Easy
0.71-0.85	Easy
0.30-0.70	Moderate
0.15-0.29	Difficult
≤0.14	Very Difficult

Item Discrimination Index is the degree to which an item can differentiate students who are knowledgeable and those who are not knowledgeable. If the discrimination index is high, it means that the item can distinguish the test taker. There are two common ways to do item discrimination index (Meyer, 2014). The first way is by computing the difference between top 27% and bottom 27% of the examinees. But, by using this way, 46% of test takers are not taken into account. Thus, giving limited information about the characteristic of the whole test takers. The second way Is by correlating item score and total test score from every individual or it is called item-total correlation. The statistics that is used to analyze the data is by using point-biserial correlation. the following is the standard used for item discrimination (Ebel & Frisbie, 1991).

Table 2. Classification for Facility Value

Facility Value	**Decision**
>0.30	Used
0.20-0.29	Revised
<0.20	Dropped

Distractors quality is regarding about how plausible is the option choices to distract not knowledgeable examinees. If the value is high, it means that the distractor is chosen by many examinees. The value of 0 indicates that there is no examinee chooses the distracters meaning that it doesn't have the ability to distract the examinees. The distractor quality is computed by using point-biserial correlation by doing distractor-total correlation. The expected value of a working distractor is negative correlation which indicates that those who are not knowledgeable will answer the question incorrectly, hence negative correlation to the total score.

Reliability is the internal consistency of the test. The range of reliability value is 0 to 1. The closer the reliability value to the maximum number, the less the measurement error in the test score, hence the more consistent the measurement. There are many kinds for reliability based on the nature of the test. In that regard, KR-21 is in line with the nature of multiple-choice test. For this analysis, KR-21 will be used for analyzing the reliability of the test.

Regarding item analysis for this project, a Windows software which is called jMetrik was used to help analyzing the data. By conducting the analysis by using jMetrik, facility value, discrimination index, distracters quality, and reliability can be explored.

Data Collection and Data Analysis Method

The population for this study was fourth semester students in English Education Department who took Reading 3 class. The sample was 24 students in class C in the fourth semester. The data were taken by giving the 40-item reading test to the students. Then, the data were recapped and analyzed by conducting Classical Test Theory analysis with the assistance of a Windows software jMetrik. By using this software, the facility value, discrimination index, distractors quality, and the reliability of the test were examined.

Findings and Discussion

The multiple-choice test has been tried out to 24 English Education Department students. The data collected were then recapped into an excel file and converted to text file. The data were analyzed by administering Classical Test Theory (CTT) which involves facility value, discrimination index, quality of distractors and reliability analyses. The analyses were computed by a computer software jMETRIK. The result of the analysis will be discussed from item level, option level, and test level.The following is the summary of the analysis.

Table 3. Summary of CTT analysis

Item	Facility Value	Discrimination Index	Distracters' Quality A	B	C	D	Decision
1	0.840	0.240	-0.330	0.240	-0.357	0.110	Revise question
2	0.880	0.486	0.486	-0.251	-0.448	-0.291	Used
3	0.840	-0.023	-0.132	-0.023	-0.063	NaN	Dropped
4	0.880	0.081	NaN	-0.207	NaN	0.081	Dropped
5	0.120	0.488	-0.414	NaN	0.488	-0.260	Used, Revise option B
6	0.560	-0.058	-0.012	-0.251	-0.058	-0.091	Dropped
7	0.800	0.193	0.193	-0.244	-0.187	-0.132	Dropped

Item	Facility Value	Discrimination Index	Distracters' Quality A	B	C	D	Decision
8	0.920	0.112	NaN	-0.216	0.112	NaN	Dropped
9	0.960	0.015	0.015	NaN	NaN	-0.092	Dropped
10	0.960	0.177	NaN	0.177	NaN	-0.251	Dropped
11	0.840	0.218	-0.159	-0.132	0.218	NaN	Revise question and option D
12	0.600	0.168	0.168	0.103	-0.135	-0.433	Dropped
13	0.760	0.273	-0.101	0.273	-0.414	-0.187	Revise question
14	0.520	0.619	-0.441	-0.130	0.619	-0.496	Used
15	0.680	0.364	0.364	-0.273	-0.316	NaN	Used, Revise option D
16	0.920	0.141	-0.052	NaN	-0.291	0.141	Revise question and option B
17	0.320	0.425	-0.410	0.103	0.425	-0.371	Used
18	0.880	0.181	NaN	-0.448	-0.043	0.181	Dropped
19	0.720	0.313	0.313	-0.330	-0.168	-0.448	Used
20	0.360	0.123	0.029	-0.205	0.123	-0.230	Dropped
21	1.000	NaN	NaN	NaN	NaN	NaN	Dropped
22	0.880	0.512	NaN	0.512	NaN	-0.601	Used, Revise option A, C
23	0.480	0.177	-0.199	NaN	0.177	-0.357	Dropped
24	0.800	0.465	-0.448	-0.291	0.465	-0.301	Used
25	0.280	0.398	0.398	-0.159	-0.458	NaN	Used, Revise option D
26	0.120	0.514	0.514	-0.247	-0.131	-0.298	Used
27	0.800	0.508	0.508	-0.330	NaN	-0.514	Used, Revise option C
28	0.120	0.258	-0.314	-0.035	-0.122	0.258	Revise question
29	0.800	0.317	-0.330	0.317	-0.291	-0.207	Used
30	0.800	0.132	-0.330	-0.087	-0.132	0.132	Dropped
31	0.520	0.238	0.068	0.074	0.238	-0.687	Revise question
32	0.800	0.380	-0.533	-0.130	NaN	0.380	Used, Revise option C
33	0.680	0.282	NaN	-0.330	-0.331	0.282	Revise question and option A
34	0.920	0.473	0.473	-0.330	NaN	-0.448	Used, Revise option C
35	0.400	-0.100	-0.039	-0.100	0.120	-0.321	Drop
36	0.720	0.239	-0.132	0.239	-0.369	NaN	Revise question and option D
37	0.920	0.321	-0.132	NaN	-0.448	0.321	Used, Revise option B
38	0.440	0.070	-0.510	0.117	-0.093	0.070	Dropped
39	0.160	-0.035	-0.507	-0.035	0.120	0.086	Dropped
40	0.320	0.575	-0.441	-0.132	0.575	-0.314	Used
KR-21 Reliability (overall40 items) = 0.664							
KR-21 Reliability (24 used and revised items) = 0.742							

Facility Value

The table above shows the facility value for every item in the test. The facility value gives an information about the difficulty level of the items. To easily consider which questions are very hard, hard, moderate, easy, and very easy, the facility value will be consulted to a standard set by Hopkins and Aten (1989). The following is the proportion.

Table 4. Proportion of Difficulty Level of the items

Difficulty	Frequency	Percentage	Item Number
Very Easy	11	27.5%	2,3,8,9,10,16,18,21,22,34,37
Easy	12	30.0%	1,3,7,11,13,19,24,27,29,30,32,36
Moderate	12	30.0%	6,12,14,15,17,20,23,31,33,35,38,40
Difficulty	2	5.0%	25,39
Very Difficult	3	7.5%	5,26,28
Average Difficulty = 0.758 (Easy)			

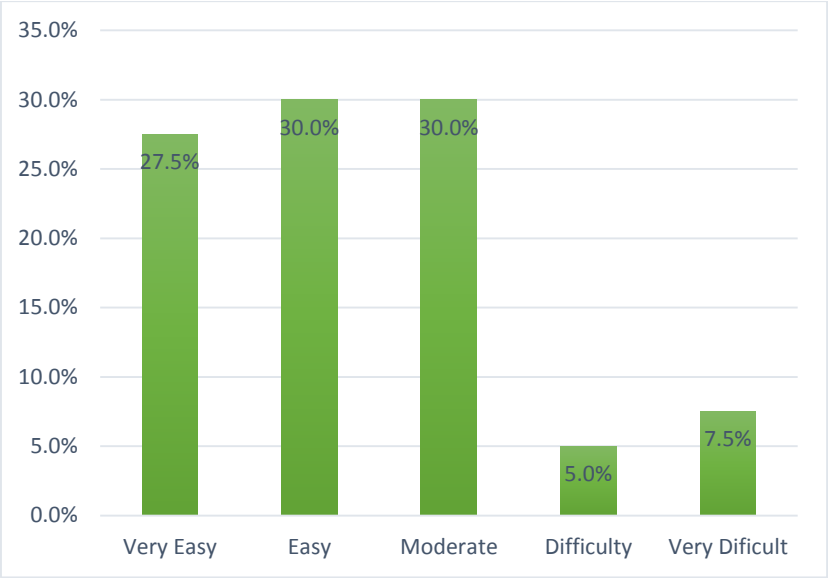

Figure 2. Proportion of Difficulty Level of the items

Based on the above table, the majority of the items have the difficulty level of very easy (11 items, 27.5%) and easy (12 items, 30%), 12 items (30%) are moderate, while hard and very hard items are the minority with 2 items (5%) and 3 items (7.5%) respectively. As seen from the barchart, the distribution of difficulty level of the items is not equal. The test seems to lean toward easy items. It is also supported by the difficulty level of the overall items, revealing the value of 0.758 which means as overal, the test was easy. On a closer examination, the reason why the test is mostly easy is because items were mainly dealing with *remembering* and *understanding* level based on Bloom's taxonomy. The example is below.

2. *In line 3, the word "bizarre" is closest meaning to*
   (A) odd
   (B) marine
   (C) simple
   (D) *Rare*

The question number 2 was assessing about the word meaning in a text provided in the test. To answer this question, the examinee doesn't need to read the text and doesn't need any critical thinking to answer the question. When he/she can recall the meaning of "bizzare", it can be answered very easily. Furthermore, the bloom's taxonomy level of this question is remembering which doesn't reflect the critical thinking intended. Another question which was very easy can also be seen from the following item.

8. *Compared with other sea creatures the sea cucumber is very*
   (A) dangerous
   (B) intelligent
   (C) strange
   (D) fat

To answer the above question, the examinee needs to read the provided text to extract the information in the text. This question has done that. However, the answer can be explicitly found in the text, so the question can be easily answered. To make students think more critically, implicit meaning-type question is expected from a question. When implicit meaning is asked in a question, students will move from remembering/understanding to more advance level (analyzing and evaluating), which generate critical thinking.
The example of The following is an example of question with moderate level of difficulty with impicit information to answer the question.

14. *What does the author imply about the United States and Canada?*
   (A) They value folk cultures
   (B) They have no social classes.

(C) They have popular cultures.
(D) They do not value individualism.

The above question shared the same nature with question number 8. The difference is a how to get information from the text. To answer the question, students need to analyze the information and extract the intended meaning from it. Unlike question number 8, the cognitive level in this question has reached analyzing cognitive level based on Bloom's Taxonomy which generate more critical thinking compared to question number 8.

5. The fourth paragraph of the passage primarily discusses
(A) the reproduction of sea cucumbers
(B) the food sources of sea cucumbers
(C) the eating habits of sea cucumbers
(D) *threats to sea cucumbers' existence*

The above question is an example of question which needs analysis to answer with very hard difficulty index. From the question, it is known that to answer it, the students need to analyze the idea of the whole paragraph and formulate the main idea of the passage. It is a good example of how a question should be made; to make students go beyond remembering and understanding to the level of analyzing the text.

Based on some samples of the question, it seems to indicate that the difficulty level of the question goes hand in hand with the cognitive level. It can be inferred that the formulation of HOTS-based test should go beyond remembering and understanding to analyzing and evaluating. Even the questions will be likely to become harder, but that is a logical consequence of developing a more challenging test for critical reading. That being said, hard question does not always mean assessing HOTS. There will be more lot consideration in confidently saying that.

Discrimination Index

The discrimination index ranges from -1 to. The criteria for discrimination index is:

a. If the discrimination index is equal or above 0.3, the item is considered a satisfying item, thus it can be used to collect data;
b. If the discrimination index is between 0.2 and 0,3, the item can be used after revision;
c. If the discrimination index is below 0.2 or NaN, the item is not working, thus it needs to be dropped.

The table above shows that the discrimination indexes are vary from -0.100 to 0.619. to better understand the proportion of used, revised, and dropped items can be seen from the table and piechart below.

Table 5. The Proportion of Used, Revised, and Dropped Items

Decision	Frequency	Percentage	Item Number
Used	16	40%	2,5,14,15,17,19,22,24,25,26,27,29,32,34,37,40
Revise	8	20%	1,11,13,16,28,31,33,36
Dropped	16	40%	3,4,6,7,8,9,10,12,18,20,21,23,30,35,38,39

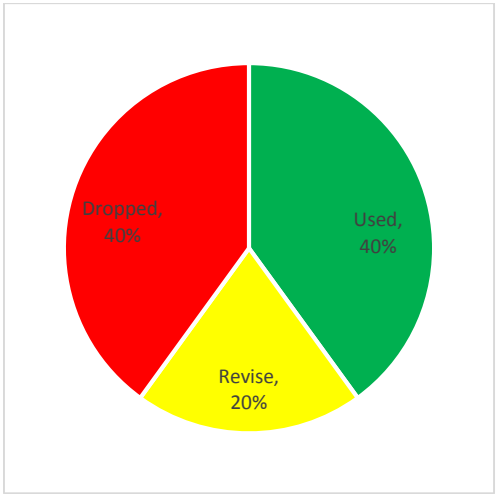

Figure 3. Proportion of Used, Revised, and Dropped Items based on Discrimination Index Analysis

The pie chart shows that 16 items (40%) could not meet the criteria, hence need to be dropped. 8 items (20%) could be used but need revision. Moreover, there are only 16 items (40%) which could be directly used without revision.

The analysis reveals that only 40% questions can be directly used. The possible factor to affect this result can be investigated through the facility value of the items. Because the questions are relatively easy, so that many students can answer correctly regardless which group they belong to (upper or lower group). The following table shows the relationship between facility value and the discrimination index.

Table 6. Proportion of Difficulty level of the Dropped Items

Item	FF	Diff	DI
3	0.84	easy	-0.023
4	0.88	very easy	0.081
6	0.56	moderate	-0.058
7	0.8	easy	0.193
8	0.92	very easy	0.112
9	0.96	very easy	0.015
10	0.96	very easy	0.177
12	0.6	moderate	0.168
18	0.88	very easy	0.181
20	0.36	moderate	0.123
21	1	very easy	NaN
23	0.48	moderate	0.177
30	0.8	easy	0.132
35	0.4	moderate	-0.1
38	0.44	moderate	0.07
39	0.16	hard	-0.035

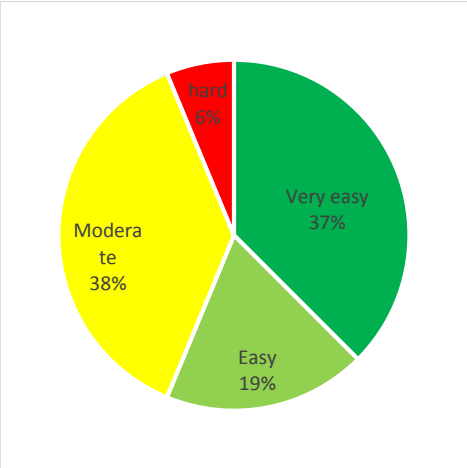

Figure 4. Proportion of Difficulty level of the Dropped Items

Based on the table and the proportion revealed by the pie chart, 56% dropped items are easy and very easy questions, while 38% and 6% are moderate and hard questions respectively. From this fact, it seems to indicate that many questions having low discrimination indexes are because of very limited information provided by the items caused by the easiness of the questions. On the other word, the questions can be answered regardless in which group the students belong to (upper or lower group), resulting in very limited information needed to run discrimination index analysis.

Example about dropped and used questions need to be examine closely to give cleared picture about the test. The following arequestion 14 and 26 which were able to discriminate students who are knowledgeable and those who are not.

14. *What does the author imply about the United States and Canada?*
    (A) They value folk cultures
    (B) They have no social classes.
    (C) They have popular cultures.
    (D) They do not value individualism

One plausible reason why the question has a good discrimination index is because the question asked for implied information from the text. Not all students can answer this kind of question because finding implied information needs more analysis process than explicit information. That is why, although the question difficulty was moderate (facility value = 0.52), it was able to differentiate between students who are knowledgeable and those who are not.

Different perspective regarding why a question can have high discrimination index can be seen from the following question number 26.

26. *The word "tolerate" in line 11 is closest in meaning to*
    (A) endure
    (B) replace
    (C) compensate
    (D) reduce

When consulted to Bloom's taxonomy, the above question seems to fit to remembering level since the students just need to recall the synonym of "tolerate".However, it becomes tricky when two answer choices "compensate" for literal meaning and "endure" for figurative sense can be related to "tolerate". That is why, the question now belongs to understanding level since the students need a context to answer the questions. High achieving students seemed to be able to answer the question correctly because they understood the context of the word, while low achieving students were distracted by the more literal word. This question is a very good example about how critical reading synonym items need to be develop; not only providing students with literal meaning of the word, but also meaning in context. By doing this, the critical sense of the students can be assessed.

The sample of questions with low discrimination indexes can be seen from question number 18 below.

18. *Which of the following would probably NOT be found in a folk culture?*
    (A) A carpenter
    (B) A farmer
    (C) A weaver
    (D) A banker

Question number 18 expected students to read the text to be able to answer it. However, the discrimination value of the item is 0.181 indicating that the question was not able to differentiate upper-lower students. A closer look into the question reveals that the flaw is related to the answer choices which were too obvious. Since folk culture is traditional, we expect that the jobs available in the culture will also be traditional, as seen from the answer choices, option D ("A Banker") is the only job which is not traditional. Thus, the question can be easily answered without examining the text, hence cannot discriminate the students.

Distractors Quality

Distractors quality is to assess how well the distractors can 'trick' not knowledgeable students into choosing incorrect choice. The standard set for a distractor to be said as working is when the distractor quality index is below 0 (having negative value). That being said, not working distractors are those with NaN value, indicating there was no student to choose the distractor. Table 3 gives full information about which distractors wereworking properly and which were not. The following table is the proportion of items with all distractors working and items with distractor(s) not working.

Table 7. Proportion of items with or without distractor problem

Criteria	Frequency	Proportion
Items with All Dictators working Properly	20	50%
Items with Distractor(s) not working Properly	20	50%

The table above indicates that the test shares equal proportion between items with all working distractors and items with distractor(s) not working. A closer look into the distractors quality reveals that the reason why they were not chosen by the students was because the text did not provide any information or at least mention anything regarding the distractors. The following is one example of this matter.

8. Compared with other sea creatures the sea cucumber is very
(A) dangerous (distractor not working)
(B) intelligent (distractor working)
(C) strange (key)
(D) fat (distractor not working)

As seen from above question, option A and D are not working, it is because there is no information in the text regarding those distractors. It happens consistently throughout the entire test. Thus, it is suggested that answer choice should touch even slightly the information provided in the text.

Reliability of the Test

Reliability is the internal consistency of the test. The range of reliability value is 0 to 1. The closer the reliability value to the maximum number, the less the measurement error in the test score, hence the more consistent the measurement. A test can be said as consistent if the reliability value is equal or more than 0.7 ($\geq$ 0.7). the reliability for this test is evaluated by KR-21 statistic as seen from Table 3.

The analysis reveals that the reliability value of the overall 40 items is 0.644 or below 0.7. It indicates that the test was not consistent and cannot be used. However, after dropping 16 dropped items, the reliability value increased to 0.742 indicating consistent test. Based on the analysis, it can be concluded that questions which were not able to discriminate students who are knowledgeable and those who are not played a significant role on the consistency of the test. Thus, they need to be immediately dropped. However, the revised items still can be used.

Conclusion and Suggestion

The conclusion than can be drawn from the analysis are:

1. In general, the test can be categorized as easy test with average difficulty level of 0.758
2. 40% of the items can be used to collect students' data, 8% need revision, and 40% has to be dropped or cannot be used to collect students' data.
3. 50% of total items have good distractors, while the other 50% have at least one distractors which was not working properly.
4. The reliability of the test including 24 items (excluding 16 dropped items) is 0.742, indicating that the test is consistent in assessing students' reading competency.

The suggestions that can be proposed are as follows:

1. It is suggested that more consideration should be made in regard to the difficulty level of the items, many questions were dropped are likely caused by the questions that are too easy, resulting in very limited information provided to run discrimination index analysis. For further test development, it should be taken into account;
2. It is suggested that more consideration should be made in deciding the level of cognitive process for the items which can be consulted to Bloom's Taxonomy. For future development of critical reading test, the items should be mostly in the level of analyze and evaluate in order to foster students' critical thinking skill.

References

Afflerbach, P., Cho, B. Y., & Kim, J. Y. (2015). Conceptualizing and assessing higher-order thinking in reading. *Theory Into Practice*, *54*(3), 203-212.

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Allyn & Bacon.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Donnelly, Christina (2014) "The Use of Case Based Multiple Choice Questions for Assessing Large Group Teaching: Implications on Student's Learning," *Irish Journal of Academic Practice*: Vol. 3: Iss. 1, Article 12.

Frisbie, D. A., & Ebel, R. L. (1991). Essentials of educational measurement.

Hassan, S., Amin, R. M., Bt. Mohd Amin Rebuan, H., & Thwe Aung, M. M. (2016). Item analysis, reliability statistics and standard error of measurement to improve the quality and impact of multiple choice questions in undergraduate medical education in Faculty of Medicine at UniSZA. *Malaysian Journal of Public Health Medicine*, *16*(3), 7-15.

Hopkins, C. D., & Antes, R. L. (1989). *Classroom testing: construction*. FE Peacock Pub.

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment. *ETS Research Report Series*, *2014*(1), 1-23.

Meyer, J. P. (2014). *Applied measurement with jMetrik*. Routledge.

Tankersley, K. (2003). *Threads of reading: strategies for literacy development*. ASCD.

Yu, J. (2015). Analysis of critical reading strategies and its effect on college English reading. *Theory and Practice in Language Studies*, *5*(1), 134.