p-ISSN 2301-7554, e-ISSN 2541-3643
https://journal.uniku.ac.id/index.php/ERJEE

ITEM ANALYSIS OF READING COMPREHENSION TEST FOR POST-GRADUATE STUDENTS

Ari Arifin Danuwijaya
English Education Department, Faculty of Language and Arts Education, Universitas Pendidikan Indonesia, Indonesia
E-mail: aridanuwijaya@upi.edu

APA Citation: Danuwijaya, A. A. (2018). Item analysis of reading comprehension test for post-graduate students. *English Review: Journal of English Education, 7*(1), 29-40. doi: 10.25134/erjee.v7i1.1493.

Received: 14-08-2018 Accepted: 23-10-2018 Published: 01-12-2018

Abstract: Developing a test is a complex and reiterative process which subject to revision even if the items were developed by skilful item writers. Many commercial test publishers need to conduct test analysis, rather than trusting the item writers' judgement and skills to improve the quality of items that need to be proven statistically after trying out was performed. This study is a part of test development process which aims to analyse the reading comprehension test items. One hundred multiple choice questions were pilot tested to 50 postgraduate students in one university. The pilot testing was aimed to investigate item quality which can further be developed better. The responses were then analysed using Classical Test Theory and using psychometric software called *Lertap*. The results showed that item difficulty level was mostly average. In terms of item discrimination, more than half of the total items were categorized marginal which required further modifications. This study suggests some recommendation that can be useful to improve the quality of the developed items.

Keywords: *reading comprehension; item analysis; classical test theory; item difficulty; test development.*

INTRODUCTION

Tests have been widely used to demonstrate level of proficiency of the students, and at the same time function as policy instruments to implement educational standards (Phakiti & Roever, 2011). In many universities, tests have become the tools used to complete the requirements in the process of admission. However, in other universities, the policies have changed in which tests play a significant role to determine not only student admission, but also graduation from their academic programs (Ma & Cheng, 2015; Mustafa & Apriadi, 2016). With the increasing demand of proficiency test for postgraduate students in Indonesian universities, many universities locally develop testing instrument that assess students' proficiency level in which most of the tests were in the form of multiple choice questions. However, studies on investigating item characteristics, such as item difficulty, item distractors, and others, in reading test are not widely exposed by the test developers or language centres in Indonesia.

Item analysis is a crucial part in a test development process as it functions to provide information about items that should be improved in terms of quality for later tests or even be eliminated due to misleading (Quaigrain & Arhin, 2007). This part is often used in the creating item banking, and its iterative nature in analysing items could help test developers to examine whether one test is a sound test both pedagogically and psychometrically and to achieve better teaching and learning (Tarrant, Ware, & Mohammed, 2009; Ananthakrishnan, 2000). For the use of English language learners, it is suggested that the characteristics of a test should be carefully reviewed and analysed (Abedi, 2002).

Several studies have been conducted to examine the processes of test development, such as item analysis in multiple choice questions in the field of education (Boopathiraj & Chellamani, 2013), medical science (Hingorjo & Jaleel, 2012; Mehta & Mokhasi, 2014; Patil, Palve, Vell, & Boratne, 2016), and social work (Qaqish, 2006); and the processes of item writing in language studies (Kim *et al.*, 2010; Spaan, 2006, 2007). Spaan's study (2006), for example, provides a practical approach of test development and item specifications. Some steps were proposed to be taken by test developers, such as test purpose writing, study analysis and construct analysis, test design, and task and item specification development. Another study conducted by Kim *et al.* (2010) recounts personal journey in the process of item writing which reveals the issues and dynamics in item writing processes. As the item analysis is hardly found in English language testing, particularly in reading, this present study aims to provide an analysis of multiple choice items in reading test.

Reading Comprehension

Reading comprehension can be defined as the ability to understand vocabulary in order to paraphrase and make a summary of information from the text (Manarin, Carey, Rathburn, & Ryland, 2015). It is the activity to reconstruct a message from written symbols to a form of a language, and it involves many cognitive processes and combines both decoding process and inferential activity so that readers can really comprehend the text (Feng & Chen, 2016; Grabe, 1997; Kendeou, Muis, & Fulton, 2011). The process is divided into two categories: lower- and higher-level processes (Grabe & Stoller, 2002; Grabe, 2009). According to Grabe (2009), lower-level processes involve word recognition, syntactic parsing and semantic-proposition encoding, while higher-level processes require text comprehension, in which good readers summarize important information from the text (Grabe & Stoller, 2002).

In the context of English as a foreign language, reading English textbooks becomes a big issue, particularly for students with non-English background. In Thailand, for example, reading skills seems to be the big problem to the students because most of them find reading English texts is difficult (Phantharakphonga & Pothithab, 2014). Within the context of university setting, reading comprehension is part of critical reading that can be a determinant to academic success. Lowes, Peters, and Turner (2004) argue that reading is essential to understand basic concepts of a subject, to gather information for completing assignments, and to improve English skill, particularly to increase vocabulary. One of many characteristics of reading at higher education is critical reading which involves such features as identifying patterns of textual elements, distinguishing main and supporting ideas, making credible evaluation and arguments, and making relevant inference about the text. Turner, Ireland, Krenus and Pointon (2011) further explains in university setting, extensive and careful reading is required in order to examine some different competing theories, for example, that leads to different ideas and information.

There are a number of skills that can be assessed in reading comprehension. Davis (1968), as cited in Alderson (2000), suggests eight reading skills, including recalling word meanings, drawing inferences about meaning of a word in context, finding answers to questions answered explicitly or in paraphrase, weaving together ideas in the content, drawing inferences from the content, recognizing a writer's purpose, identifying a writer's technique, and following the structure of a passage.

Reading Comprehension Test (RCT)

A test is a tool that serves to make decisions related curriculum and other areas (Brown, 1995; Carr, 2011; Spaan, 2006). Brown (2004) points out that a test is a way of measuring one's ability, knowledge, or performance in a given domain. In language testing context, most tests measure test takers' competence, such ability to perform language skills to speak, write, listen, or read to one subset of language. These performance-based tests sample the test-takers' actual use of language which infers general competence. A test of reading comprehension, for example, may consist of several short reading passages each followed by a limited number of comprehension questions. From the results of the test, the examiner may infer a certain level of general reading ability (Brown, 2004).

Reading Comprehension Test (RCT) is an instrument that measures university students' abilities in reading a wide array of texts. This high-stakes test is developed to assess reading skills of postgraduate school students, in which its result can be used as the requirement for students to have thesis examination. Students need to obtain a certain score to have the examination. If the score cannot be reached, students are not allowed to take the exam. Thus, RCT can be viewed as a high-stakes test in the university. It consists of 100 multiple choice questions which test some skills in reading, such as the ability to understand main information in the text, scan detailed information, get the meaning of words, understand pronoun reference question, make inferences from the text, identify not-explicitly-stated information, and locate information in the text.

Constructing a test is not a simple task. It involves a science and art of many complex tasks, such as planning, test preparation and administration, scoring, statistical analysis, and test result report (Brown, 2004; Downing, 2010). One of crucial stages in test development is statistical analysis of a test. Statistics are beneficial in language testing. During the first trial of the items, statistics can inform an analysis of each tested item. For example, a test designer can take the advantage of statistics to examine if the distractors work well in a multiple choice or the item is too difficult to answer. In addition, according to Kunnan and Carr (2013), statistical analysis functions to provide a summary of test takers' performance in a form of test scores. It informs the test developer about the descriptive statistics of the test takers' performance, such as the average score (mean), the most occurring score (mode), and the overall variation from the average (standard deviation value) (Kunnan & Carr, 2013).

Classical Test Theory

One of essential statistical tools in the analysis of language test is Classical Test Theory (CTT). CTT has given a significant contribution to the area of language testing. According to Brown (2012), many university courses and textbooks in language testing discuss the general idea of CTT. Besides, most language teachers and practitioners use CTT in their practice in language testing. Thus, CTT appears to serve as the foundation for understanding all aspects of language testing, and understanding CTT becomes vital for a language test designer because CTT is a precondition for comprehending and using more forms of analysis (Brown, 2012).

Brown (2012) suggests that there are main methods in CTT including item analysis (item facility, item discrimination, and distractor efficiency analysis), reliability estimates, the standard error of measurement, and various validity analysis. Item analysis is a crucial procedure to improve the quality of objective test by investigating how effective an item is. The result of the analysis informs which item needs to be included, modified, or eliminated in the test. There are three procedures in test analysis: item facility, item discrimination, and distractor efficiency.

Item facility (IF), often called as item difficulty, describes the proportion of test takers who correctly answered the item (Brown, 2012; Carr, 2011). Brown (2012) argues that if 95% of the test takers answer an item correctly, then the item is categorized as very easy; on the other hand, an item is viewed as very difficult if it is answered by 11% of the test takers. In addition, Carr (2011) suggests that the values of item facility range from 0.0 to 1.0 indicating none of the students answered correctly and every test taker answered it right respectively. Ideal items would be items of intermediate facility – the items that

30-70% of the test takers answered correctly or within the range of 0.3 to 0.7 (Brown, 2012; Carr, 2011). Items with IF below 0.30 are usually deemed to be too difficult, and items with IF above 0.7 are considered too easy (Carr, 2011).

In addition to item facility, an item can be analysed in terms of how well the given item distinguish between test takers with high and low ability, or commonly called as item discrimination (ID) (Carr, 2011; Thorndike & Thorndike-Christ, 2010). There are two ways of calculating discrimination. One way is by "subtracting the number of students who got the item correct in the lower group (NL) from the number who got it correct in the upper group (NU) and dividing the difference by the number of the group (N)" (Thorndike & Thorndike-Christ, 2010, p. 308). Another way to estimate item discrimination is using a correlational approach, in which a correlation coefficient is calculated between the item score and the total score, known as point-biserial correlation coefficient (Brown, 2012; Carr, 2011). With the existence of scoring machine or psychometric software, the value of correlation computation can be easily performed. The value of ID ranges between 0.00 and 1.00 and it can be positive or negative. Items with the highest values (more than 0.5) need to be retained, the ones with the lowest (below 0.2) need to be eliminated, and the ones between 0.2 and 0.5 should be consider for modification (Thorndike & Thorndike-Christ, 2010). According to Brown (2012), it is desirable to include the items with high discrimination indexes in the revised test because including the high discrimination index items will lead to more reliable measurement overall, whereas including items with low discrimination will lead to less reliable measurement.

The last item analysis is distractor efficiency analysis or distractor analysis. According to Brown (2012), the distractors are essential parts of an item and function to show a relationship between the total test score and the distractor chosen. Low scoring students should choose the distractors more often while students with high scores choose the correct option. Thus, the function of efficiency analysis is to investigate how efficient the distractors are to divert test takers from the correct answer (Brown, 2012). According to Quaigrain and Arhin (2017), if there is a distractor chosen by less than five per cent of the test takers, the distractor is called as a non-functioning distractor (NFD). By analysing the distractors, it is easier for test developer to make a decision whether the distractors are revised, replaced, or removed.

The rationale for the development of the reading test is the needs to construct up-to-date language test aiming to investigate postgraduate English reading skills in the university. The main objective of the study is to investigate item difficulty, discrimination, and distractor efficiency of multiple choice test items in reading comprehension.

METHOD

This study aimed to examine the process of test development, particularly in the process of analysing multiple choice questions in reading skills, to improve item quality. The study was conducted at one language centre in one public university in Indonesia. The centre was chosen because its availability to provide items for try-outs and analysis. As this is a trial test, only a small subset of target population involved in this study to provide useful information about the items (Spaan, 2007). Fifty postgraduate students from different majors (educational management, science and mathematics education, social sciences, and non-formal education) who aged from 20 – 45 years were involved in this study. The students were invited to take part in the pilot testing of the items in December 2016.

A hundred of multiple choice questions had been written in 2016 but not yet pilot tested to get the evidence on item quality. These questions were written to provide information about reading proficiency among postgraduate students. The items were developed by five English language teachers having more than three years of teaching experience. The items were constructed based on ten reading passages with the topics ranging from education, literature, social sciences, and others. The questions for each passage, as shown in Table 1, aim at assessing reading skills, including skills in skimming for main idea, scanning for stated detailed information, deducing meaning of unfamiliar words, pronoun resolution, making inference, understanding unstated information, and scanning to locate specific information (Alderson, 2000; Shirvan, 2016).

Table 1. *Reading skills in multiple choice questions*

Reading Skills	**Number of Questions**	**Item Number**
Skimming for main idea	6	1,11,21,51,61,90
Scanning for detailed information	29	2,4,6,9,15,16,19,27,28,32,39,47,50,55,57,62,65,66,68,74,77, 80,84,85,86,89 ,96,97,100
Deducing the meaning and the use of unfamiliar lexical item	32	3,8,10,13,14,18,23,24,26,31,36,37,45,48,49,53,56,58, 59,63,64,67,70,72,73,75,78,82,87,93,94,98
Pronoun resolution	10	5,25,29,33,35,42,46,54,71,83
Making inference	10	7,20,22,30,41,43,69,76,81,88
Understanding information when not explicitly stated	7	12,34,38,44,91,92,99
Scanning to locate specific information	6	17,40,52,60,79,95,
Total Questions	100	

The items were one-correct answer type, having a stem and four options, one of them being correct and the other three being ‘distractors’. The test takers were required to select the correct choice and fill the answer on a separate answer sheet. Each correct response was awarded 1 mark. No mark was given for blank response or incorrect answer. There was no negative marking. The maximum possible score was 100 and the minimum 0.

Based on students’ responses, the test items were then analysed using Laboratory of Educational Research Test Analysis Package (Lertap) psychometric software (Nelson, 2001). Lertap is a computer program to process and analyse results from tests and surveys. With the use of Microsoft Excel interface, the program and manual are user-friendly making teachers, instructors, and researchers easy to perform classical item, reliability, and dependability analyses of raw test or survey data (Carr, 2004). The program can also be used to score and perform reliability analysis for both affective and cognitive subtests. Each question was analysed in terms of its level of difficulty, which was measured by the difficulty index (p-value), power of discrimination, and distractor analysis. The cut-off values for item difficulty used Carr’s (2011) guideline with three categories: easy, average, and hard level of difficulty with difficulty index of less than 0.3, between 0.3 and 0.7, and more than 0.7 respectively. For item discrimination index, there are four categories of items based on its discrimination index: poor (DI < 0.15), marginal (0.15 < DI < 0.24), good (0.25 < DI < 0.34), and excellent (> 0.35) (Hingorjo & Jaleel, 2012).

RESULTS AND DISCUSSION

In the development of test instrument, there are some crucial steps that should be conducted to ensure the high quality of newly developed items. One of the crucial steps is to conduct pilot testing, or also called as pre-testing or trialling. Carr (2011) argues that pilot testing is vital to ensure that the constructed items produce responses that are expected from the test takers. Besides, pilot testing helps test developer to estimate the reliability of the test and to examine whether each item is appropriately written and, in multiple choice items, the distractors can work well to discriminate lower and higher ability of the test takers. Thus, item can be analysed for further improvement.

After the pilot testing was administered to the respondents, the item responses were analysed by using statistical descriptive. The results showed that the scores of 50 test takers ranged from 23 to 68 with the mean score of 47.06 and the standard deviation was 10.05. The median score was 47 and the inter-quartile range value was 45. The skewness and kurtosis values for the scores were -0.15 and -0.18, respectively. As the values between -2 and +2, it is acceptable to prove normal univariate distribution. The test takers were divided into three groups. The mean scores of lower, middle, and upper groups were 35.6 (SD = 5.3), 47.3 (SD = 2.8), and 58.2 (SD = 4.8), respectively.

Based on the analysis, the Cronbach's alpha index to measure the test reliability was 0.80. The value was categorized as a large alpha value indicating "that the items are tapping a common domain" (Wells & Wollack, 2003, p. 4). However, for a test that functions as a high-stake standardized test, it has a certain criteria in test reliability. Wells and Wollack (2003, p. 5) argue that standardized tests should have higher reliability coefficient as the test is given only once and function "to draw conclusions about each student's level on the trait of interest". They further suggest that the internal consistency coefficients in a high-stake standardized test should be at least 0.9, and for lower-stake test should have at least 0.80 or 0.85 (Wells & Wollack, 2003).

Item analysis

One-hundred multiple choice items were processed using Lertap. Another result shows a brief statistical report as illustrated in a plot of item difficulty by discrimination in Figure 1. The plot suggests that even though the test was high in reliability (coefficient alpha = 0.80) with some items were found to have average difficulty (ranging from 0.3 to 0.7), more than half of the items were categorized marginal to poor (62%) even having minus discrimination index.

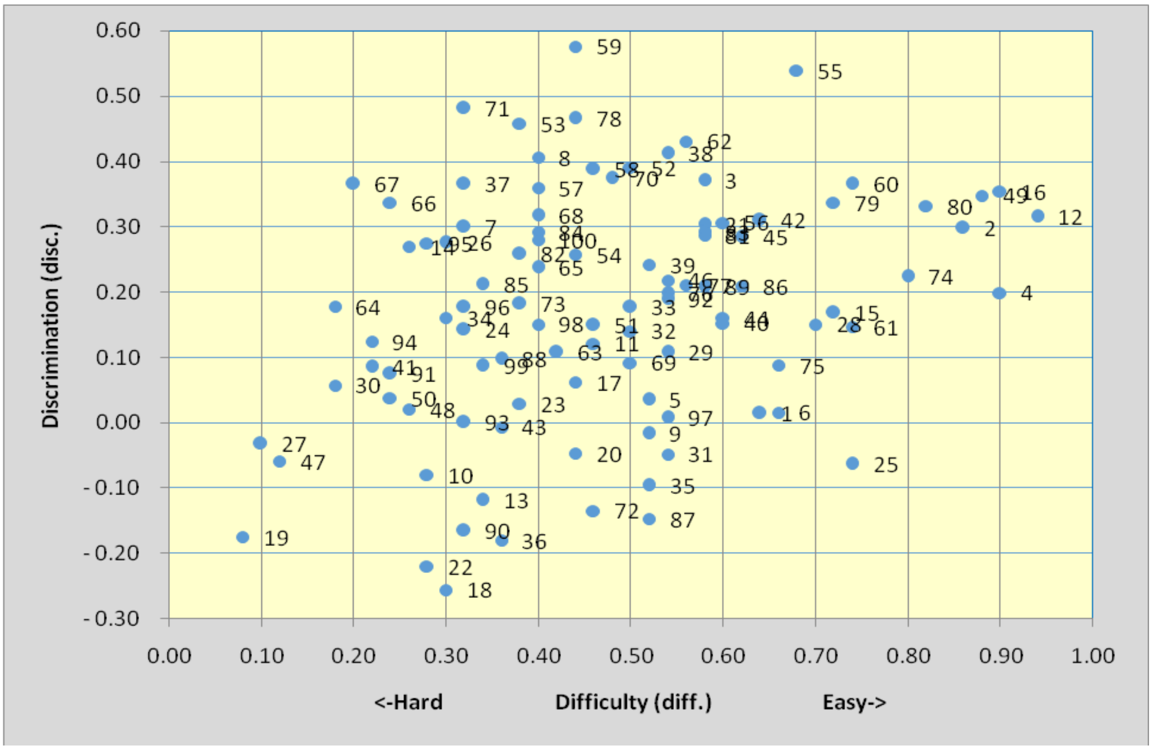

Figure 1. *The plot lot of items based on item difficulty and discrimination index*

Item difficulty

The results of analysis showed that the average value of item difficulty was 0.47 and standard deviation of 0.19. Figure 2 illustrates the proportion of the items based on the level of difficulty: easy, average, and hard. It is found that majority of items (71%) were of average difficulty. Some items were outside the desired range of 0.3 to 0.7. Out of 100 items, there were 16 items below 0.3 indicating more difficult items and seven items above 0.7 indicating easier items. Hingorjo and Jaleel (2012) suggest that items with average level of difficulty is more desirable, items with easy category can be placed at the beginning of the test as ‘warm up’ questions, and difficult items should be reviewed for language confusion or even incorrect key.

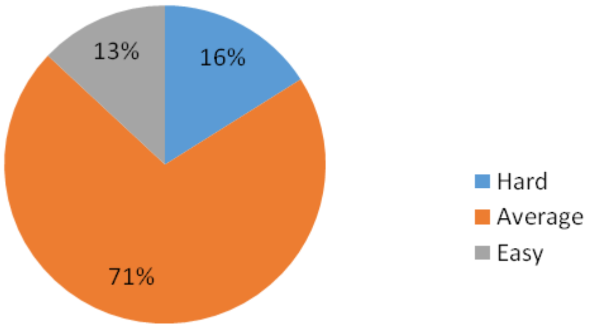

Figure 2. *The proportion of items based on item difficulty level*

Table 2 below shows the item classification based on the reading skills. Items with easy level were mostly dominated by questions related to scanning for detailed question skills. For items with average level of difficulty, items related to deducing meaning and the use of unfamiliar lexical items had the highest proportion accounting for 24%, followed by questions related to scanning for detailed questions (18%). Meanwhile, items with high level of difficulty were dominated by items related to deducing meaning and the use of unfamiliar lexical items (6%), scanning detailed information (5%), and making inference (3%).

Table 2. *Classification of items based on reading skills*

Reading Skills	**Item Difficulty Level**		
	Easy	**Average**	**Hard**
Skimming for main idea	1	5	
Scanning for detailed information	6	18	5
Deducing the meaning and the use of unfamiliar lexical item	2	24	6
Pronoun resolution	1	9	
Making inference		7	3
Understanding information when not explicitly stated	1	5	1
Scanning to locate specific information	2	3	1
Total Questions	**13**	**71**	**16**

Item discrimination

Item discrimination has a significant role to examine if an item is of low or high quality. Items that function well to discriminate between students with different abilities are desirable and will increase reliability (Nelson, 2001; Wells & Wollack, 2003). According to Nelson (2001), discrimination index can be calculated in two approaches, called as the correlation and upper-lower method. Measuring item discrimination with correlation approach is known as point-biserial correlation or pb(r), which is the correlation between students’ scores on the item and the student's overall score. The analysis of items using Lertap displays a full statistics that consist of point-biserial correlation, as shown in Table 3.

Table 3. *Samples of the statistical analysis on test items*

1 (c3)

option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	1,00	32	0,64	0,01	0,02	47,53	0,05	
B	0,00	18	0,36	-0,06	-0,08	46,22	-0,08	
C	0,00	0	0,00	0,00	0,00	0,00	0,00	<-no
D	0,00	0	0,00	0,00	0,00	0,00	0,00	<-no

2 (c4)

option	wt.	n	p	pb(r)	b(r)	avg.	z	
A	0,00	0	0,00	0,00	0,00	0,00	0,00	<-no
B	0,00	3	0,06	-0,19	-0,39	39,33	-0,77	
C	1,00	43	0,86	0,30	0,46	48,40	0,13	
D	0,00	4	0,08	-0,25	-0,46	38,50	-0,85	

Table 3 shows the samples of item statistical analysis that has pb(r) value in two items (Question 1 and Question 2). The values of point-biserial in Question 1 and Question 2 were 0.01 and 0.03, respectively. According to Wells and Wollack (2003), a large positive pb(r) shows that test takers with higher scores tended to answer the item correctly, while lower score test takers responded incorrectly. In addition, item with small positive pb(r) does not significantly improved the test reliability, but even it can cause to reduce the reliability in some cases. In contrast, item with negative pb(r) will reduce test reliability, and it is preferable that an item has pb(r) more than 0.20 (Wells & Wollack, 2003). To conclude, Question 1 has a low positive pb(r) and Question 2 has a desirable pb(r) that can improve reliability.

On the other hand, the index of item discrimination also describes the ability of an item to discriminate test takers with high and low scores. According to Hingorjo and Jaleel (2012), there are four categories of items based on its discrimination index: poor (DI < 0.15), marginal (0.15 < DI < 0.24), good (0.25 < DI < 0.34), and excellent (> 0.35). Figure 3 shows the percentage of item classification based on discrimination index. It is found that more than half of the total items had poor and marginal discrimination index, accounting for 39% and 23%, respectively. Meanwhile, there were only 20% of good items and 18% of excellent items found in the test.

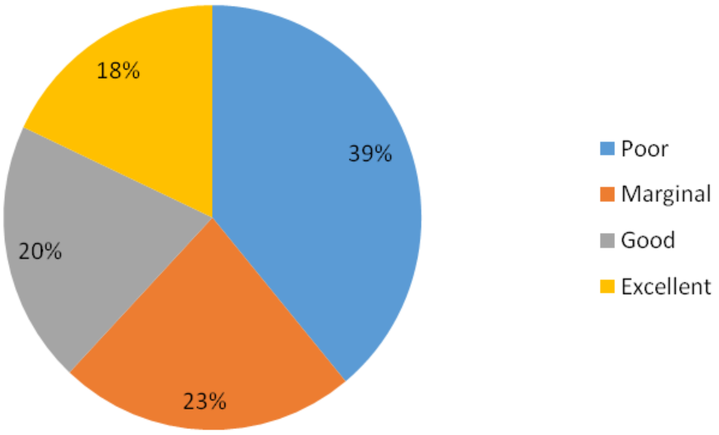

Figure 3. *The proportion of items based on discrimination index*

Table 4 shows the proportion of items based on reading skills. Based on the table, there were 14 items related to deducing the meaning and the use of familiar lexical items that were categorized as poor items, followed by scanning detailed information (18 items). For marginal level, most items were related to scanning for detailed questions. Items related to scanning for detailed information and deducing the meaning and the use of familiar lexical items were mostly found in the category of good and excellent.

Table 4. *Classification of items based on reading skills*

Reading Skills	**Item Difficulty Level**			
	Poor	**Marginal**	**Good**	**Excellent**
Skimming for main idea	3	2	1	0
Scanning for detailed information	8	11	6	4
Deducing the meaning and the use of unfamiliar lexical item	14	3	5	10
Pronoun resolution	4	2	3	1
Making inference	7	1	2	0
Understanding information when not explicitly stated	2	3	1	1
Scanning to locate specific information	1	1	2	2
Total Questions	**39**	**23**	**20**	**18**

Distractor analysis

According to Hingorjo and Jaleel (2012), distractor analysis is essential to examine whether the distractors function well – low scoring students chose the distractor more, compared to higher scoring students. With the analysis, it makes possible for the test developer to revise, replace, and even remove the distractors.

The test analysed consists of 100 questions with four options each; thus, the total number of distractors were 300. The analysis found that 39 out of 300 (13%) of the distractors were categorized as non-functioning distractors. Non-functioning distractors were defined as distractors that were chosen by less than five per cent of the test takers or even those which were not selected at all by the test takers (Hingorjo & Jaleel, 2012). A distractor was categorized as working distractor when it was chosen, or some of lower examinees chose it. However, a distractor which was not chosen by anyone or fooling higher ability examinees does not function well (Qaqish, 2006). This type of distractor is not contributing to test ability to discriminate the good students from the poor students, and thus it should be replaced or eliminated (Kehoe, 1995). Table 5 shows some samples of items with non-functioning distractors.

Table 5. *Samples of questions with non-functioning distractors*

Q1			**Q2**			**Q4**		
Option	n	/50	Option	N	/50	Option		
A	32	64.0%	B	3	6.0%	A	4	8.0%
B	18	36.0%	C	43	86.0%	C	45	90.0%
			D	4	8.0%	D	1	2.0%

For Question 1, 50 test takers answered the question, and 32 of the test takers chose the right answer (option A). This question is in average difficulty level. However, the rest of the test takers (n = 18) chose B, leaving option C and D being not chosen by anyone. In this case, option C and D failed to function as good distractors. Options C and D were categorized as non-functioning distractors. Another sample for Question 2, 43 out of 50 test takers answered the correct answer (option B). Option B and D worked well as distractors because the options were chosen by some test takers. However, no one answered option A, and thus, it is called as the non-functioning distractor. This case was similar to Question 4, in which option B and D were the non-functioning distractors.

According to Carr (2011), some problematic items can be improved by revision, and the items that require revision are those having negative item point-biserials, particularly items with large magnitudes. Based on the analysis, there were 17 items out of 100 which had negative item point-of biserials, two of which had large magnitude. An item with negative point-biserial and high magnitude can be exemplified by Question 22 (-0.22). This reading comprehension item had item difficulty of 0.28 and a point-biserial of -0.22, which was problematic. The item was in difficult category and had negative discrimination. The question can be seen in Figure 4, and the passage was about the Incan Empire.

1. It can be inferred that Pharaoh . . .
   A. referred to the name of an empire in Egypt
   B. possessed powerful supremacy in Egypt
   C. was an Egyptian god
   D. lived for hundreds of years

Figure 4. *Example of an item with problematic distractors*

Table 6 shows the distractor analysis of the example item. Based on the responses of the item, it can be concluded that the distractors functioned well, as all three distractors were answered by the test takers. However, it can be seen that item C was more attractive that item B (the correct answer). The important detail of the response was shown in the value of point-biserial. Any distractors should have negative point-biserial coefficient, indicating that test takers who chose a wrong answer tended to have lower scores and vice versa (Carr, 2011). Based on Table 5, option B, the correct answer, has a high negative point-biserial, which was highly problematic. It indicated that the option was more attractive to low-ability test takers than high-ability test takers. In contrast, option C was attractive to higher-ability test takers. Therefore, the item required further changes, particularly the options. If the answer is B, then option C should be modified.

Table 6. *Distractor analysis results for the example item*

Option	n	p	pb(r)
A	7	0,14	-0,34
B	14	0,28	-0,22
C	25	0,50	0,51
D	4	0,08	-0,21

CONCLUSION

Item analysis has provided useful information about the characteristics of items in one test. Some items, after the analysis, might be revised, changed, or even removed. Based on the analysis above, it is found that the test had high reliability with Cronbach's alpha coefficient of 0.8. As the nature of the test was high-stakes which function to make decision of the graduation for postgraduate school students, the reliability of the test needed to be improved, and one of them was by improving item quality. Items should function to discriminate between students with different abilities. Based on the findings, many items were categorized as marginal and poor category in terms of discrimination index. Thus, these items should be treated for better development by either modification or deletion from the test set. Based on the difficulty level, most of the items (71%) were categorized in the average level of difficulty, which was desirable for a test. The analysis showed that item with very easy and very difficult level needed further treatment.

Based on the above findings, there are some recommendations that can be put into consideration for future development. Constructing high-stakes test takes about one to two years to complete and involves staff from many different capabilities, such as experts in subject matter, test specialist, editors, psychmetricians, and many others. As a result to this, the first recommendation to improve the quality of item development is to ensure that items were written based on the test purpose and by experienced test writer. Writing good items required knowledge and practice, and thus it is essential to have experienced colleagues share the process of item writing. Second, it is quite appropriate to provide training materials for teachers to learn how to write better items. Thorndike and Thorndike-Christ (2010) suggest that writing good items is a learnable skill and there are some principles of making an item. However, the construction of the questions is not simple as it requires certain skills and it has rules of writing the item. Thus, it would be beneficial to provide training to ensure that the items written can discriminate properly and all the distractors function well.

REFERENCES

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment*, *8*(3), 231-257.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Ananthakrishnan, N. (2000). Item analysis-validation and banking of MCQs. In N. Ananthkrishnan, K. R. Sethuraman, & S. Kumar, Medical education principles and practice. *Pondichery: JIPMER*, 131-137.

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International Journal of Social Science & Interdisciplinary Research, 2*(2), 189–193.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Pearson Education.

Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston: Heinle and Heinle Publishers.

Brown, J. D. (2012). Classical test theory. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing measurement*. Accessed on: March 23, 2017 from http://doi.org/10.4324/9780203181287.ch22.

Carr, N. (2004). A review of Lertap (Laboratory of Educational Research Test Analysis Package) 5.2. *International Journal of Testing*, *4*(2), 189–195.

Carr, N. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Downing, S. M. (2010). Test development. In *International Encyclopedia of Education* (3rd ed., pp. 159–165). Elsevier.

Feng, Q., & Chen, L. (2016). A study on teaching methods of reading comprehension strategies by comparison between tem-4 reading comprehension and IELTS academic reading comprehension. *Journal of Language Teaching and Research*, *7*(6), 1174–1180.

Grabe, W. (1997). Current developments in second language reading research. *TESOL Quarterly*, *25*(3), 375–460.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Harlow: Longman.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index, and distractor efficiency. *The Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research, and Evaluation, 4*(10).

Kendeou, P., Muis, K. R., & Fulton, S. (2011). Reader and text factors in reading comprehension processes. *Journal of Research in Reading, 34*(4), 365–383.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullin, V. (2010). A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writers' characteristics. *Language Assessment Quarterly*, *7*(2), 160–174. doi: 10.1080/15434300903473989.

Kunnan, A. J., & Carr, N. T. (2013). Statistical analysis of test results. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–7). Oxford: Blackwell Publishing.

Lowes, R., Peters, H., & Turner, M. (2004). *The international student's guide: Studying in English at university*. Thousand Oaks, CA: SAGE Publications.

Ma, J., & Cheng, L. (2015). Chinese students' perceptions of the value of test preparation courses for the TOEFL iBT. *TESL Canada Journal, 33*(1), 58–79.

Manarin, K., Carey, M., Rathburn, M., & Ryland, G. (2015). *Critical reading in higher education: Academic goals and social engagement*. Bloomington, Indiana: Indiana University Press.

Mehta, G., & Mokhasi, V. (2014). Item analysis of multiple choice questions - An assessment of the assessment tool. *International Journal of*

Health Sciences and Research, *4*(7), 197–202.

Mustafa, F., & Apriadi, H. (2016). DIY: Designing a reading test as reliable as a paper-based TOEFL designed by ETS. In *Proceedings of the 1st English Education International Conference* (pp. 402–407). Banda Aceh.

Nelson, L. R. (2001). *Item analysis for test and surveys using Lertap 5*. Perth: Curtin University of Technology.

Patil, R., Palve, S. B., Vell, K., & Boratne, A. V. (2016). Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *International Journal of Community Medicine and Public Health*, *3*(6), 1612–1616.

Phakiti, A., & Roever, C. (2011). Current issues and trends in language assessment in Australia and New Zealand. *Language Assessment Quarterly*, *8*(2), 103–107.

Phantharakphonga, P., & Pothithab, S. (2014). Development of English reading comprehension by using concept maps. *Procedia - Social and Behavioral Sciences*, *116*, 497–501.

Qaqish, B. (2006). Developing multiple choice tests for social work trainings. In B. Johnson, M. Henderson, & M. Thibedeau (Eds.), *Eighth Annual National Human Services Training Evaluation Symposium* (pp. 91–111). Berkeley, California: California Social Work Education Center, University of California.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *12*, 1–11. doi: 10.1080/2331186X.2017.1301013.

Shirvan, M. E. (2016). Assessing and improving general English university students' main sub-skills of reading compression: A case of University of Bojnord. *Sino-US English Teaching*, *13*(4), 245–260. doi: 10.17265/1539-8072/2016.04.002.

Spaan, M. (2006). Test and item specifications development. *Language Assessment Quarterly*, *3*(1), 71–79. doi: 10.1207/s15434311laq0301.

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly*, *4*(3), 279–293. doi: 10.1080/15434300701462937.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education, 9*(1), 40.

Thorndike, R. M., & Thorndike-Christ, T. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Boston: Pearson Education.

Turner, K., Ireland, L., Krenus, B., & Pointon, L. (2011). *Essential academic skills*. Melbourne: Oxford University Press.

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. Madison: University of Wisconsin.