CONSTRUCTING ENGLISH READING COMPREHENSION ABILITY TEST FOR HIGH SCHOOL STUDENTS BY USING AN IRT MODEL

Wint Wah Wah Tun[1], San Win[2]

Abstract

Although the mastery of reading comprehension in English is becoming more important day by day, a large number of students in Myanmar seem to suffer from poor reading comprehension. Appropriate assessment is the essential to assess the English reading comprehension ability of students so as to provide them with necessary feedbacks and instructions. So, the main aim of this study is to construct the standardized English Reading Comprehension Ability Test (ERCAT) for Myanmar High School students. In this study, descriptive survey design and quantitative approach was used. Snow and the RAND Reading Group's (2002) Framework of Reading Comprehension was used as the primary framework for constructing the test in this study. Barrett's Taxonomy of Reading Comprehension (Barrett, 1976) was used to define the cognitive demand level of the tasks included in the ERCAT. Initial item pool of the test consists of 74 items related to 25 reading passages of various lengths. Then, the pilot testing was conducted with 316 high school students from B.E.H.S Intagaw, Bago. Some items were revised and eliminated according to IRT calibration. The final version of the ERCAT included 40 items related to 15 reading passages. Then, as the field testing, the data was collected from 1026 high school students by using stratified random sampling technique. The results revealed that all items' difficulties and discriminations were between the acceptable ranges. The empirical reliability of the ERCAT is 0.9354 and the maximum test information of the test I (θ) was 19 at $\theta = 0.25$. The test had smaller standard error across the ability scale from -1.25 to +1.75 and thus, the test was informative for the average ability range as it was predicted.

Keywords: Test Construction, Reading Comprehension Ability, IRT Model, High School Students

Introduction

Reading comprehension has started to receive a special focus in foreign language teaching since it can be considered as the most important skill for getting information, and exploring and broadening academic knowledge (Azeroual, 2013). Above of all, English reading comprehension is increasingly important in creating a successful society because English is the language of information in many professional areas including science, technology, medicine, etc. It is also the main language of books, academic conferences, international business, diplomacy and sport.

Cain and Oakhill (2006) found out in their research that children who showed superior comprehension ability would later have greater educational success and their finding highlighted the importance of acquiring good comprehension skills. However, students in general encounter challenges in reading with meaning (National Reading Panel, 2000). The percent of the struggling readers in the high school context have reached about 70% in the USA. The problem of reading comprehension is a world-wide educational issue and especially, when reading in English as a foreign language, the challenge of comprehension gets worse.

At the same time, in Myanmar, most students are focusing on learning the vocabulary by heart, finding the main ideas and retrieving facts with shallow understanding of the content while they read in English. Current assessment system for high school level almost emphasizes on the memorization of the contents prescribed in the English textbook rather than the acquisition of reading comprehension ability.

[1] Assistant Lecturer, Department of Educational Psychology, Yangon University of Education

[2] Professor, Dr, Department of Educational Psychology, Yangon University of Education

Due to the lack of standardized English reading comprehension tests in Myanmar, most high school students seem to encounter difficulty in assessing their comprehension ability. Moreover, the most popular reading tests such as TOEFL junior, etc., are unaffordable to most Myanmar high school students. The present study intended to fill this gap by constructing standardized English reading ability test for Myanmar high school students. The results of this study will hopefully help to provide the high school teachers with the knowledge of constructing a standardized test using an IRT model.

Purposes of the Study

The main aim of the present study is

- ▪ to construct the English Reading Ability Test for high school students by using an IRT model

The specific objectives are

- ➢ to construct the English Reading Ability Test for students with average ability, and
- ➢ to differentiate the qualities of the items included in the test.

Definitions of Key Terms

Test Construction. Test construction refers to the science and art of planning, preparing, administering, scoring, statistically analyzing, and reporting results of test (Downing, 2010).

Reading Comprehension Ability. Reading comprehension ability is the ability to utilize lower order reading processes (including decoding and vocabulary knowledge) and higher order reading processes (including relation of text content to schema and conscious controllable processing) to understand concepts and ideas from text (Pressley, 2000).

IRT Model. An Item Response Model shows the relationship between the ability or trait measured by the instrument and an item response (DeMars, 2010).

Related Literature Review

Reading Comprehension Process

Reading is a complex cognitive skill involving many sub-skills and processes ranging from basic lower-level visual processes involved in decoding the print to higher level skills involving skills of text comprehension and higher-level thinking. According to the comprehensive theory of the reading process presented by Grabe and Stoller (2002), both lower-level decoding skills and the higher-level comprehension interact to produce successful comprehension of texts.

The lower-level processes are mostly automatic and involve linguistic skills, such as vocabulary for lexical access, grammar for syntactic parsing, and the combination of meaning and structural information for semantic proposition formation. Higher-level processes involve the coordination of ideas from a text in order to form a meaningful representation of the text. Background knowledge and reading strategies are also used for the interpretation of the text. These higher-level processes interact in more complex ways and require higher cognitive abilities (Grabe & Stoller, 2002).

Using an IRT Model in Test Construction

Whenever a test is constructed, development of test specifications, item writing, field testing the items and calibrating the items should be done first. Then, the good items are selected for the final test (Baker, 2001). A test can generally be constructed by using one of the two test theories, namely, classical test theory (CTT) and item response theory (IRT).

The major advantage of IRT models over classical test theory is that classical item and test characteristics (statistics) vary depending upon the group of candidates but IRT item and test

characteristics do not. Classical indices of item difficulty, point-biserial correlation and reliability may all change if candidate differ in ability distribution (Hathaway, Houser, & Kingsbury, 1985). An IRT model compares the difficulty of the item with the ability of the candidate and estimates the probability that the candidate will correctly answer the item. If an IRT model is used in constructing a test, a measure of the precision of the item calibration and ability estimate is available for each item and each candidate and thus, it can provide additional valuable statistical information.

Method

Sampling

The participants for this study were chosen by using stratified random sampling technique. Firstly, two states (Kachin and Rakhine) and three regions: (Mandalay, Ayeyarwaddy, and Bago) (30% of total states and regions) were selected. And then, 2 high schools under Department of Basic Education from each selected state and region were chosen. Therefore, altogether 10 high schools were chosen and then nearly 100 high school students from each high school were selected. Finally, 1026 students were chosen for this study.

Research Method

In this study, descriptive survey design and quantitative approach were used.

Research Instrumentation

Theoretical Framework Used for Constructing Reading Comprehension Ability Test

In present study, the framework used in constructing English reading comprehension ability test for high school students was Snow and the RAND Reading Group's (2002) Framework of Reading Comprehension which defined reading comprehension as the joint outcome of three combined sources of influence: the reader, the text and the activity or task for reading.

Reader: A reader brings a number of reader factors to reading, which can include motivation, prior knowledge, and other cognitive abilities.

Text: Reading necessarily requires material for the reader to read. In an assessment, that material which may be a piece of text or a set of texts related to a particular task must include sufficient information for a proficient reader to engage in meaningful comprehension and resolve the problem posed by the task.

Tasks: Task factors include the potential time and other practical constraints, the goals of the task, and the complexity or number of tasks to be completed.

A reading comprehension test is for enabling to measure the comprehension ability of students by manipulating task and text factors. Based on this framework, the selection process of text passages to be included in the present study was carried out. The researcher adapted the reading passages from Preparation Book for TOEFL Junior Test (Reading Comprehension Advanced) and The B level (B1 and B2) Exam in English (Student's Book). The consideration was taken to select the text passages with appropriate context and difficulty for the Myanmar high school students.

To ensure a broad coverage of reading comprehension domain, a broad range of text types including descriptive, narrative and expository texts were used in this study. According to the length of the text, there were 25 reading passages of different lengths of which, specifically, were 20 single and short passages with the mean length of 41 words, 4 medium passages with the mean length of 159 words, and a long passage of 5 paragraphs with the length of 411 words.

Then, a table of specifications was developed in order to construct the items with high content validity. Initial item pool of the test consists of 74 items. Many educationists and

researchers use different taxonomies to categorize the levels of comprehension tasks (Umalusi, 2011). In this study, Barrett's taxonomy of reading comprehension was used to define the intended cognitive demand level of each item.

Taxonomy Used to Define the Cognitive Demand Level of Reading Tasks

Barrett's Taxonomy of Reading Comprehension, as shown in Table 1, was used to define the cognitive demand levels of various types of reading comprehension questions in this study.

Table 1. Barrett's Taxonomy of Reading Comprehension (Barrett, 1976)

Type of cognitive demand	What questions require of the student/examinee
1.Literal Comprehension Recognition or recall of - details - main ideas - a sequence - comparison - cause and effect relationships - character traits	To locate or identify any kind of explicitly stated fact or detail (for example, names of characters or, places, likeness and differences, reasons for actions) in a reading selection/text/ material
2. Reorganization - classifying - outlining - summarizing - synthesizing	To organize, sort into categories, paraphrase or consolidate explicitly stated information or ideas in a reading selection/text/material
3.Inferential Comprehension - main ideas - supporting details - sequence - comparisons - cause and effect relationships - character traits - predicting outcomes -interpreting figurative language	To use conjecture, personal intuition, experience, background knowledge, or clues in a reading selection/text/ material as a basis of forming hypotheses and inferring details or ideas (for example, the significance of a theme, the motivation or nature of a character) which are not explicitly stated in the reading selection/ text/ material
4. Evaluation – Judgment of - reality or fantasy - fact or opinion - adequacy or validity - appropriateness -worth, desirability and acceptability	To make evaluative judgment (for example, on qualities of accuracy, acceptability, desirability, worth or probability) by comparing information or ideas presented in a reading selection/text/material using external criteria provided (by other sources/authorities) or internal criteria (students' own values, experiences, or background knowledge of the subject)
5. Appreciation -Emotional response to content -Identification with characters -Reactions to author's language use -Imagery	To show emotional and aesthetic/ literary sensitivity to the reading selection/text/material and show a reaction to the worth of its psychological and artistic elements (including literary techniques, forms, styles, and structuring)

Source: Barrett, 1976

Scoring Rubrics for English Reading Comprehension Ability Test

To assess the student's responses on various types of reading tasks, scoring rubric was developed as a scoring guide. Each item with multiple-choice format was worth one point. The tasks which require the students to select a summary or a title for each passage in a given text were in the form of matching items and all these items were worth one point per item. Each item included in a cloze procedure was also worth one point. It was worth one point for each true or false items. For the reorganization item, the students had to reconstruct the reading passage of 8 lines in correct sequential order. If half of the given sentences (4 lines) and above can be correctly reorganized, it was worth one point and if not, it would be zero.

Then, for each open-ended extended response, each student's answer must be classified with respect to the coverage of the important parts present within the answer. Depending on whether the students can identify and locate the required information in given piece of text, or not, each response was worth one point or zero. This scoring rubric was designed according to the Snow and RAND's (2002) framework of reading comprehension.

After designing the test items and planning the scoring procedures, expert review was conducted for face validity and construct validity by 18 experts in the field of English, methodology, educational psychology, and educational test and measurement from Yangon University, University of Distance Education, Yangon University of Education, Sagaing University of Education. According to the suggestions of experts, revisions in item length and wording of items were made. Next, the test was administered to 316 Grade 10 students during 2019-2020 AY. Some items were revised and eliminated according to IRT calibration. Then, the test was administered to 1241 high school students across Myanmar for the field testing. The data obtained were analyzed by using two-parameter logistic model (2PLM) of IRT with the application of BILOG-MG 3 software. Firstly, in order to use item response theory (IRT), the assumptions of unidimensionality, local independence, and model data fitness were checked.

Data Analysis and Research Findings

Checking the Assumption of Unidimensionality

The assumption of unidimensionality means that the model has single θ for each examinee (DeMars, 2010). In order to determine whether the assumption of unidimensionality was met, eigenvalue plot of the inter-item correlation matrix for the test was studied. Figure 1 showed the scree plot of the eigenvalues for all items included in the test.
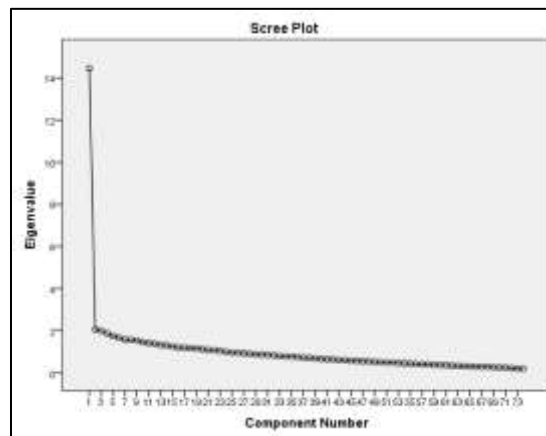

Figure 1. Scree Plot of the Eigenvalues for All Items of the Test

The dominance of the first factor can be observed in Figure1. The largest eigenvalue of the correlation matrix for all items in the test is over three times larger than the second largest eigenvalue. Thus, it could be assumed that it has the unidimensionality. When the assumption of unidimensionality is met, so is the assumption of local independence (Lord, 1980).

Checking the Assumption of Model Data Fitness

In order to check the assumption of model data fitness, the obtained data were firstly analyzed by all three popular models. According to sample size, one-parameter and two-parameter models could be used with the obtained data but three-parameter model was not fit with the observed data.
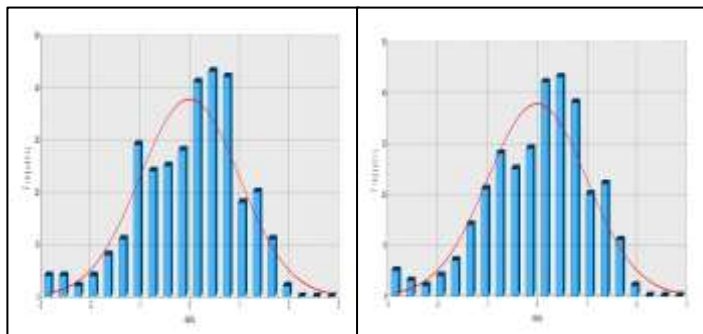
Note: ▯ = observed proportion, ——— = expected proportion

Figure 2. Observed and Expected Proportion of the Test by Appling One-Parameter Logistic Model and Two-Parameter Logistic Model

Then, goodness of the Model Data Fit was checked by Lord's Chi-square method. According to the Chi-square results, it was found that the number of model data fit items were 59 items when one-parameter logistic model was applied. Analyzing with two-parameter logistic model, the number of model data fit items were 72 items. Thus, it could be clearly observed that the number of modal data fit item was larger by using two-parameter logistic model. In addition, the difference between observed and expected proportion of the test by utilizing one-parameter and two-parameter models can be found in Figure 2.

As it can be seen in Figure 2, the residuals were greater while utilizing the one-parameter logistic model (1PLM). It could be assumed that many of the items in the test would not fit with 1PLM. The observed proportion correct was closed to the model expectation when the two-parameter logistic model (2PLM) was applied. Therefore, 2PLM is the best fit with the data of the test.

Analyzing Item Parameters of the Items in English Reading Comprehension Ability Test

Most items in the test had good discrimination and some items had moderate discrimination. But, the item no. 15, 17, 20, 22, 41, 47, 53 and 58 has discrimination less than 0.4 and so these items should not be included in the operational test. Most of the items in the test had the difficulties between the range of -2 to +2 with the exception of item no 15, 17, 22, 41, 47, 53, 58, 68, and 74. The item difficulty values of these items were so high that they were considered to be too difficult for the intended population.

Revising the Table of Specification for the ERCAT

Based on the results from IRT calibration, a table of specifications was reconstructed for the final selection of the items to be included in the test in order to ensure the content coverage of the test. Revised table of specifications for the ERCAT was described in Table 2 as follow.

Table 2. Table of Specifications for English Reading Comprehension Ability Test

Content	Literal				Inferential			Evaluative		Total Items	Total Marks
	MCQ	Re	T/F	S.Q	MCQ	T/F	Mat	MCQ	S.Q		
Reading Single & short passages (11 Passages)	4 (1)	-	-	-	5 (1)	-	-	2 (1)	-	11	11
King Bayintnaung (Narrative)	-	1 (1)	-	-	-	-	-	-	-	1	1
School Announcement (Description)	2 (1)	-	-	-	2 (1)	-	-	-	-	4	4
Daily Diet (Expository)	-	-	2 (1)	-	1 (1)	-	5 (1)	-	-	8	8
Reading Medium Passages (1+2+3)	2	1	2	-	3	-	5	-	-	13	13
Reading Long Passage The Great Fire of London (Mixed)	2 (1)	-	5 (1)	2 (1)	1 (1)	-	3(1)	1 (1)	2(1)	16	16
Total Items	**8**	**1**	**7**	**2**	**9**	**-**	**8**	**3**	**2**	**40**	
Total Marks	**18**				**17**			**5**			**40**

Note: MCQ = Multiple Choice Questions,
      Re   = Rearrange,
      T/F = True or False,
      SQ   = Short Questions
      The numbers in parentheses represent the marks allocated to each item.

Final Selection of Items for the ERCAT

In order to select the items for the operational test, the purpose of the test was defined first. The ERCAT was intended to measure the average ability students (i.e., $\theta = -2 \sim +2$). Therefore, a greater number of middle-difficulty items were selected to be able to measure most of the examinees precisely (i.e., $\theta = -2 \sim +2$). Moreover, items with fair and high discrimination were also selected so as to distinguish between low and high performing students. Item parameters of selected items included in the ERCAT were shown in Table 3.

Table 3. Item Parameters of the Selected Items in the ERCAT

Item	Item Response Theory		Item	Item Response Theory	
	a	*b*		*a*	*b*
Item 1	1.532	0.255	**Item 21**	0.835	-0.21
Item 2	1.172	0.197	**Item 22**	0.891	-0.286
Item 3	0.946	-0.022	**Item 23**	0.757	-0.042

Item	Item Response Theory		Item	Item Response Theory	
	a	*b*		*a*	*b*
Item 4	0.762	-0.041	Item 24	0.713	-0.143
Item 5	0.636	0.73	Item 25	0.564	0.104
Item 6	0.632	0.294	Item 26	0.606	0.391
Item 7	0.589	0.382	Item 27	0.944	0.024
Item 8	0.578	-0.115	Item 28	1.245	-0.292
Item 9	0.513	0.57	Item 29	1.053	-0.207
Item 10	0.603	0.103	Item 30	1.061	0.031
Item 11	0.546	-0.287	Item 31	0.794	-0.338
Item 12	0.662	0.652	Item 32	0.517	0.174
Item 13	0.748	-0.096	Item 33	0.899	0.417
Item 14	0.732	0.223	Item 34	0.962	0.598
Item 15	0.746	0.156	Item 35	0.793	0.662
Item 16	0.705	0.461	Item 36	0.718	0.769
Item 17	0.849	-0.094	Item 37	0.78	0.336
Item 18	0.735	0.659	Item 38	0.612	0.73
Item 19	1.314	0.63	Item 39	1.255	1.143
Item 20	0.995	0.65	Item 40	1.021	0.492

According to the Table 3, it was found that the item parameters of the selected items to be included in the ERCAT were heterogeneous, i.e., the item difficulty (*b*) values ranged from easy to difficult level (-0.338 to + 1.143). The item discrimination (*a*) values of all items in the test ranged from 0.513 to 1.532.

Item Characteristic Curves (ICCs) for the ERCAT

The item characteristic curves of all items in the ERCAT were graphed in Figure 3 in order to present the probability of choosing the correct answer to an item as a function of the level of the attribute being measured by the test. According to the figure 3, it can be seen clearly that the higher the student's ability level, the greater the probability of getting the correct answer. Then, it was also observed that most item characteristic curves in the test were almost parallel which is meant that most items of the test had similar discrimination. Moreover, the item characteristic curves were placed according to their difficulty, i.e., from easy to difficult, and they were closed together due to small variance in difficulty.
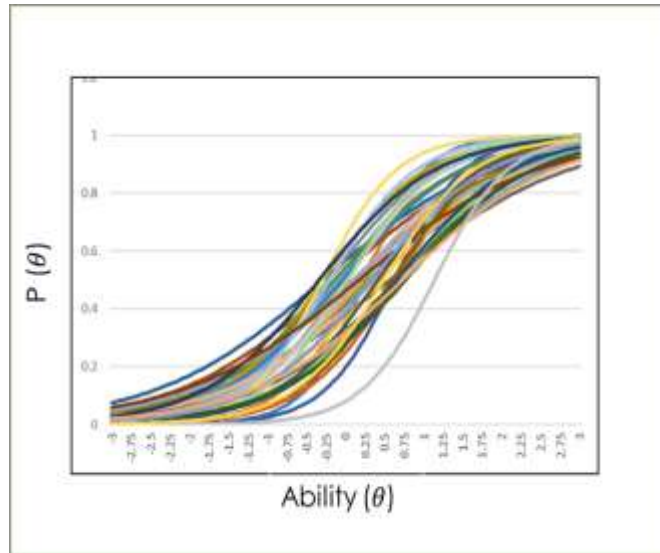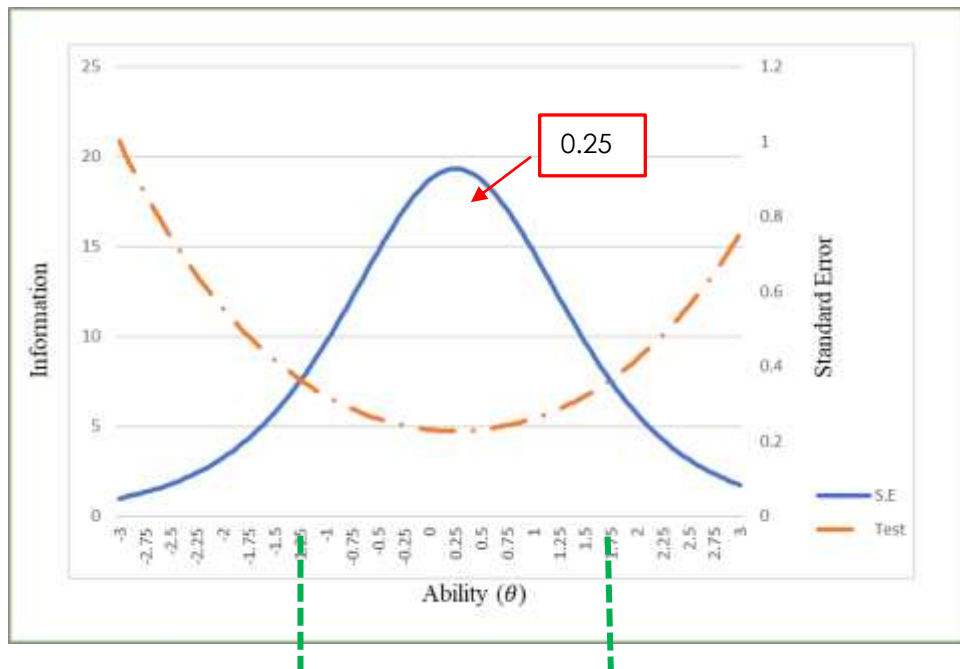
Figure 3. Item Characteristic Curves of All Items in the ERCAT

Test Information Function (TIF) of the ERCAT

In order to know the maximum amount of information obtained by the ERCAT precisely, the test information function of the ERCAT was illustrated in Figure 4.

Figure 4. Test Information Function (TIF) for the ERCAT

As shown in Figure 4, the maximum information of the test I (θ) was 19 at $\theta = 0.25$. In other words, this test gave the maximum information about the students who had ability +0.25 The test had smaller standard error across the ability scale from -1.25 to +1.75. Thus, it can be concluded that the test was informative for the average ability range as it was predicted. In addition, it could be interpreted that the ERCAT could be used to determine the student group whose average ability is $\theta = 0$ ($-2 < \theta < +2$).

Discussion and Recommendations

In this study, English Reading Comprehension Ability Test (ERCAT) for Myanmar high school students was constructed by applying the two-parameter logistic model (2PLM) of IRT. Initially, the 74-item test with 25 reading passages was relatively difficult and so, it was restored as the item pool. New optimal developed test was reconstructed with 40 items which were related to 15 reading passages for the target test information function. The developed test would provide enough information for high school students with average English reading comprehension ability. By using the ERCAT, the educators, teachers and test administers can assess and predict the general English reading comprehension ability of high school students.

There are some research limitations and several extended studies that can be further undertaken. Firstly, according to test information function curve, it was found that the test composed of 40 items could be suitable for high school students whose reading comprehension ability range is from -1.25 to +1.75, but it cannot discriminate well for the high school students with higher ability levels (above $\theta = +2$) and lower ability levels (below $\theta = -1.5$). It is still necessary to fill more easy items and to arrange them from the difficult items to easy items across the ability scale, until the test information function range -3 to +3 is achieved.

Secondly, the number of 1026 students from 10 high schools were selected from two states (Kachin and Rakhine) and three regions (Mandalay, Ayeyarwaddy, and Bago) to participate in this study. It will be necessary to replicate with larger sample size from different regions and states. Moreover, further studies should be carried out to construct more items that can be added to the item bank for assessing reading comprehension ability of high school students. Finally, the test developed in this study can only be administered to the high school students and so it cannot provide any evidence on the reading comprehension ability of students with different education levels. Thus, it was calling for the further researches to develop similar tests that can be administered to students of different education levels.

Conclusion

In order to estimate the quality of education and the achievement of the students, various types of tests are used in educational fields. Teachers can get valuable information about how well their students are learning from the test results. Besides providing feedback to their students, the test results can also help the teacher to be able to reflect themselves on their own teaching methods and curriculum materials. Therefore, systematically constructed tests that can provide reliable evaluation or results are the essential core of the assessment system and so as to deliver the quality education to students.

Acknowledgements

References

Azeroual, D. (2013). *Investigating the reading difficulties of Algerians EST students with regard to their general English knowledge*. Arab World English Journal.

Baker, F. B. (2001). *The Basic of Item Response Theory* (2nd ed.). United States: ERIC Clearinghouse on Assessment and Evaluation.

Barrett, T. C. (1976). *Taxonomy of reading comprehension. In Teaching reading in the middle grades*. Eds. R. Smith and T. C. Barrett. Reading. MA.: Addison-Wesley.

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities, 44*(5), 431-443.

DeMars, C. (2010). *Item Response Theory: Understanding Statistics Measurement*. Oxford University Press, Oxford. Retrieved June 12, 2018, from http://doi.org/10.1093/acprof:oso/9780195377033

Downing, S.M. (2010). *International Encyclopedia of Education* (3rd ed). McGaw, Australia.

Grabe, W. & Stoller, F. (2002). *Teaching and researching reading.* London: Pearson Education.

Hathaway, W., Houser, R., & Kingsbury, G. (1985*). A regional and local item response theory-based test item bank system*. (ERIC Document Reproduction No. ED 284 883).

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

National Reading Panel. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (Tech. Rep. No. NIH Publication No. 00-4769). Washington, DC: National Institute of Child Health and Human Development.

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M.L. Kamil, P.B. Mosenthal, P.D. Person and R. Barr (Eds.), *Handbook of reading research, Volume 3*, pp. 545-561. Mahwah, Nj:Erlbaum.

Snow, C. and the RAND Corporation. (2002). *Reading for Understanding: Toward an R and D Program in Reading Comprehension*. RAND Reading Study Group, Santa Monica, CA. Retrieved February 20, 2018, from http://www.rand.org/

Umalusi. (2011). Comparative Analysis of the National Senior Certificate Home Language, 2008-2010. Afrikaans; English; isiNdebele; isiXhosa; isiZulu; and Siswati. B. Lepota and N.L. Thwala. Pretoria: Umalusi.