CONSTRUCTING A COMPUTER-BASED ENGLISH TEST FOR MIDDLE SCHOOL STUDENTS

Wint Wah Tun[1] and Nu Nu Khaing [2]

Abstract

The main purpose of this study was to construct a Computer-based English test for Middle School Students. To achieve the purpose of the study, three alternate forms of Grade 9 English achievement tests were first constructed according to the table of specifications. Each test contained 50 multiple-choice items. The sample were randomly selected from Yangon Region by using Quantitative survey research design. The sample size of totally 1514 students from eleven high schools participated in this study. According to the results, it can be assumed that the tests were neither too easy nor too difficult for students and they also discriminated well among examinees with different abilities. In order to compare the students' results, norm-tables were constructed by transforming raw scores to scaled scores. The norm tables can help the teachers to interpret the achievement levels of the students, even the students administered different forms of tests. For computer-based test, Quiz Faber computer software was used. To create an online computer-based test, Google Forms application was applied and QR code was used to send the test in online.

Keywords: computer-based tests (CBT), item response theory (IRT), scaling

Introduction

Importance of the Study

Assessment plays a critical role in the field of education, allowing teachers as well as administrators to make important decisions regarding the proficiency, placement, and achievement of students. The most common assessment tools in the education system all over the world is the "test". They are a useful and essential part of teaching and learning. They play an important role in today's schools and other aspects of life. At all levels of education (i.e., kindergarten through graduate), most professional certificating procedures and many employment opportunities place a high reliance on test performance. So tests need to be qualify and to be ensure fairness for all test takers. Moreover, they need to give valid, reliable and useful information concerning student achievement.

In educational assessment, paper-and-pencil tests (PPT) and computer-based tests (CBT) are being used with considerable success for measuring degrees of student achievement. With the increase availability of computers, many assessments are being administered as computer-based tests (CBTs). Computerized exams frequently are perceived as being "state of art" or automatically better than traditional, standardized test, paper-and-pencil exams (Bridgeman, 2009). CBTs provide several advantages over paper-and-pencil tests including ease and flexibility of administering and grading tests, as well as, allowing for the development of technology-based environment. These benefits have made CBT increasingly popular.

Moreover, the types of questions asked via traditional paper-based assessments or exams can also be asked via a computer-assisted assessment or exam. CBTs can be administered and scored more accurately, quickly and securely than paper-and-pencil tests. CBTs can either be an online test or already saved (downloaded) in the computer. With online test, the students can take the test according to their convenience from any location by using Internet. And the results can

[1] Senior Assistant Teacher, Integrated High School for Sports and Education, Tarmwe, Yangon Region
[2] Dr, Lecturer, Department of Educational Psychology, Yangon University of Education

be given quickly and accurately after the examination in both types of computer-based tests. This can give beneficial features to the students and teachers.

Computer-based testing can provide many advantages which can improve educational assessment. They can increase the depth of student knowledge and skill that can be assessed, improve the range of accuracy for test results, increase the efficiency of the assessment process and improve the fairness of testing. Therefore, computer-based tests are used successfully in university admissions, placement, certification and licensure testing.

In testing situation where alternate forms are used, it is not meaningful to compare the examinee's test results by only raw scores. A number of correct raw score of 20 on one test form does not necessarily indicate the same level of achievement as a number-correct raw score of 20 on another test form. Moreover, raw scores cannot actually represent the level of different students even on the same test. As a consequence, information contained in a raw score is limited. Almost all large assessments report scale scores to provide information that cannot be reflected in a raw score. So, in order to compare, explain and make proper decisions, the test users construct norm tables by transforming raw scores to scaled scores. Norm tables can help teachers to interpret the achievement levels of students and the students can know their relative standing in the group.

Recently, much research has been focused in developing and expanding the class of item response theory models to solve a wide variety of measurement problems. According to Hambleton, Swaminathan and Rogers (1991), applications of IRT include test development, item banking, differential item functioning, adaptive testing, test equating, and test scaling. A major appeal of IRT is that it provides an integrated psychometric framework for developing and scoring tests. Moreover, item response models are particularly suitable for computer-adaptive test which is one kind of computer-based test because it is possible to obtain ability estimates that are independent of the particular set of test items administered. Even though each examinee receives a different set of item, differing in difficulty, item response theory provides a framework for comparing the ability estimates of different examinees. Much of this research has focused on dealing the circumstances under which the theoretical advantages of IRT are fulfilled in practice.

In Myanmar, computer-based tests have been applied in educational assessment. IELTS (International English Language Testing System) and TOEFL (Testing of English as a Foreign Language) are examples of computer-based tests used in Myanmar. But there is very limited amount of computer-based tests used in academic subjects. Therefore, this study tried to develop the three forms of Grade 9 English test by applying Item Response Theory model and to construct a Computer-based English Test for Middle School Students.

Purposes of the Study

The purposes of the study are:

1. To construct three alternate forms of Grade 9 English achievement test by applying an IRT model;
2. To develop norm tables for three alternative tests and
3. To construct a Computer-based Test for Grade 9 English.

Method

Sample of the Study

Participants in this study were approximately 1514 Grade 9 students from Yangon region. They were divided into three groups (Group 1- 514 students, Group 2– 516 students, Group 3 – 514 students). Eleven Basic Education High Schools were selected and this study was geographically restricted to Yangon region.

Test Construction Procedures

In this study, three alternate forms of Grade 9 English Achievement test for Computer-Based Test were prepared based on the same table of specifications. Initially, item pools of Grade 9 English were constructed to investigate the qualities of items. Next, the total 200 multiple choice items were constructed with the same content and learning outcomes.

After constructing the item pools, experts' review was conducted for face validity and content validity by twelve experts from Department of Educational Psychology from Yangon University of Education. For pilot testing, the total 200 items were divided into 100 items for Form A and 100 items for Form B. Then, pilot testing was administered to the sample of 50 students from selected school in Yangon Region within two days.

According to the pilot result, some incorrect or ambiguous items were corrected and some were eliminated. Finally, the three test forms that contain 50 multiple choice items were constructed for field testing. The three test forms were constructed with the same content and same table of specifications.

Data Collection and Scoring Procedures

After constructing Form A, Form B and Form C with the same contents and same table of specifications, they were administered to 1514 Grade 9 students during the last week of November and the first week of December, 2018. A spiraling process was used to randomly assign the forms. And then, the responses of students were dichotomously scored. The correct answer for each item was given one point and the incorrect answer was scored zero point.

Software for Constructing the Computer-based Tests

In this study, QuizFaber computer software developed by Luca Galli was chosen for constructing a computer-based English test for Grade 9 student. QuizFaber is a Freeware software for windows that enables to create multimedia quizzes as HTML documents. The quiz is ready to be published on Internet, in a local network or on a local PC. It is possible to create and manage many types of questions: questions with multiple choice, questions with multiple answer, true or false questions, questions with open answer, gap filling exercises and matching words. It can be fully customize for the choice of background images, colors, sounds and font types. The quiz result can be saved on a web server, send through email, stored on the Google cloud (Google Drive) or into internet server.

Moreover, Google Forms application was chosen to construct an online computer-based test. Google Forms is a web-based application used to create forms for data collection purposes and for test administration. Students and teachers can apply Google Forms to make surveys, quizzes, tests or even registration sheets. It can be used to ask both open-ended and close-ended questions (Text, Paragraph Text, Multiple Choice, Checkboxes, etc.). The form is web-based and can be shared with respondents by sending a link, emailing a message, or embedding it into a

web page or blog post. In this study, QR code was applied to share a computer-based test in online although there are many ways to share the test form. The online test form developed in Google Forms was embedded in QR code by creating in QR Code Generator Application. When students take the test, they are given QR code of the question paper. And they can use mobile devices or computers to scan this QR code. Students can answer the question on their screen and send the results to the server.

Data Analysis and Findings

Checking the Assumption of Unidimensionality

In this study, analysis of the eigenvalues of the inter-item correlation matrix was applied to check the unidimensionality assumptions. According to the scree plots, the largest eigenvalues of Form A, Form B and Form C were about six times larger than second eigenvalues of these forms. Therefore, the result showed that the three forms held the assumption of unidimenisionality. The scree plots of Form A, Form B and Form C are shown in Figure 1, Figure 2 and Figure 3.

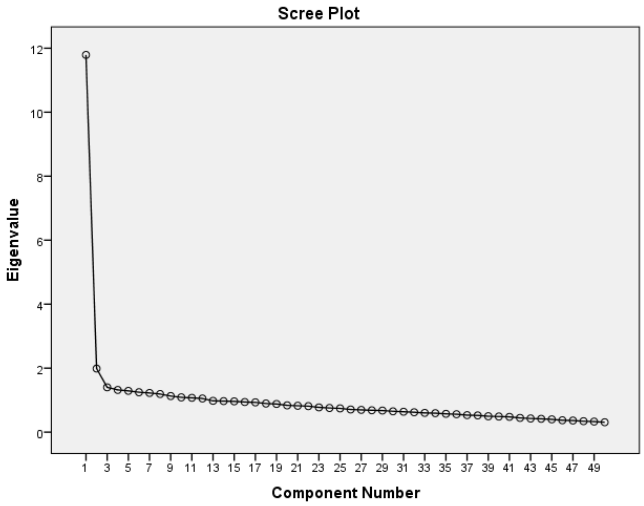

Figure 1 Scree Plot of Form A

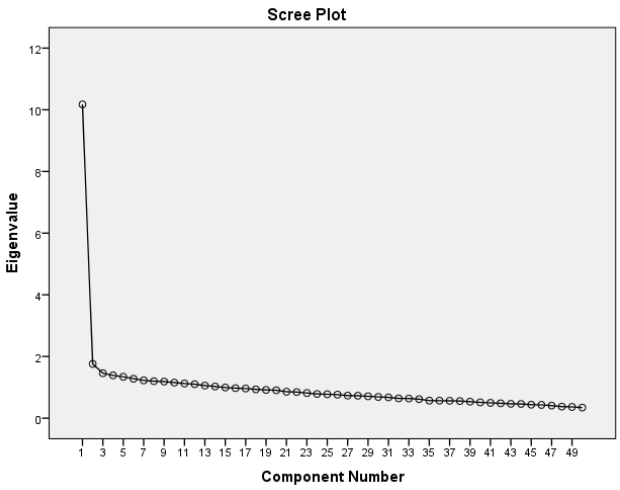

Figure 2 Scree Plot of Form B

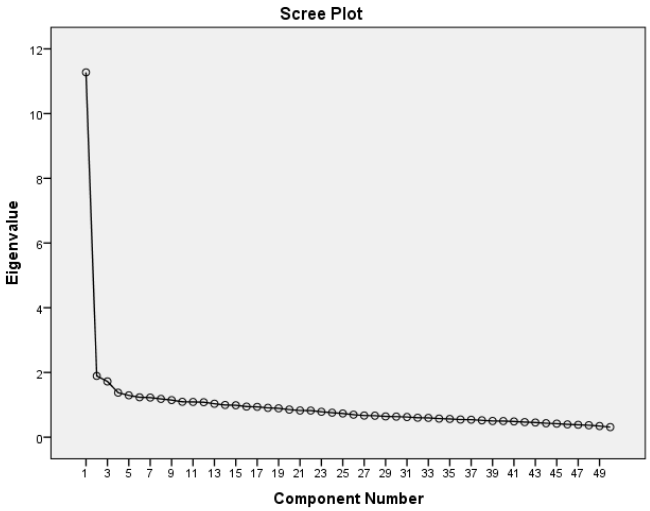

Figure 3 Scree Plot of Form C

Item Analysis by Item Response Theory

Table 1 Item Parameters (IRT) for the Three Tests

Test	Parameters					
	Discrimination (*a*)			Difficulty (*b*)		
	Mean	SD	Range	Mean	SD	Range
Form A	0.88	0.43	0.14 ~ 1.83	-0.28	0.72	-1.34 ~ +2.38
Form B	0.73	0.29	0.12 ~ 1.39	-0.30	0.86	-1.35 ~ +2.81
Form C	0.81	0.35	0.12 ~ 1.78	-0.45	0.70	-1.74 ~ +1.73

According to the Table 1, it was found that the three test forms discriminated well among the examinees with different abilities and they were neither too easy nor too difficult.

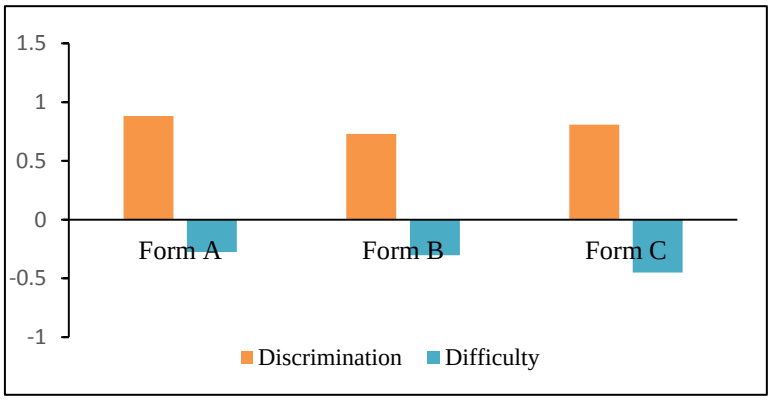

Figure 4 Comparison of Difficulty and Discrimination of the Three Tests

Comparisons of TCCs and TICs of the Three Test Forms

Test characteristics curve (TCCs) and test information curves (TICs) of three test forms were plotted and they were shown in the following figures.

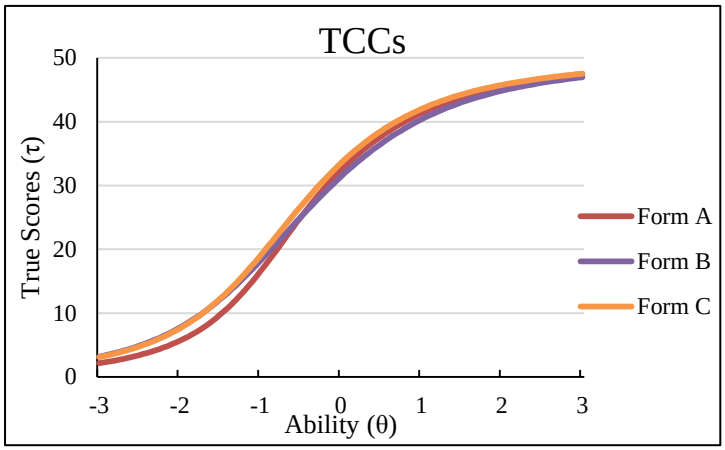

Figure 5 Comparison of Test Characteristics Curves of Three Test Forms

According to figure 5, the three test forms had appropriate difficulty and appropriate discrimination. Since the curves were parallel, the three test forms can be used alternatively.

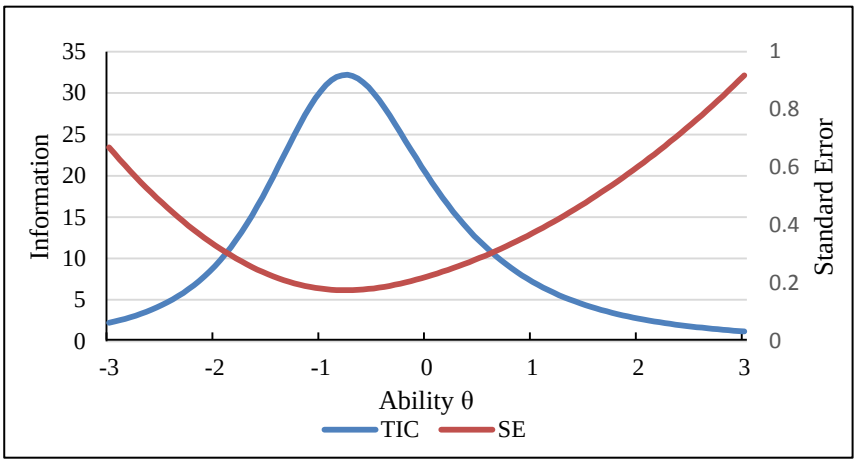

Figure 6 Test Information Curve of Form A

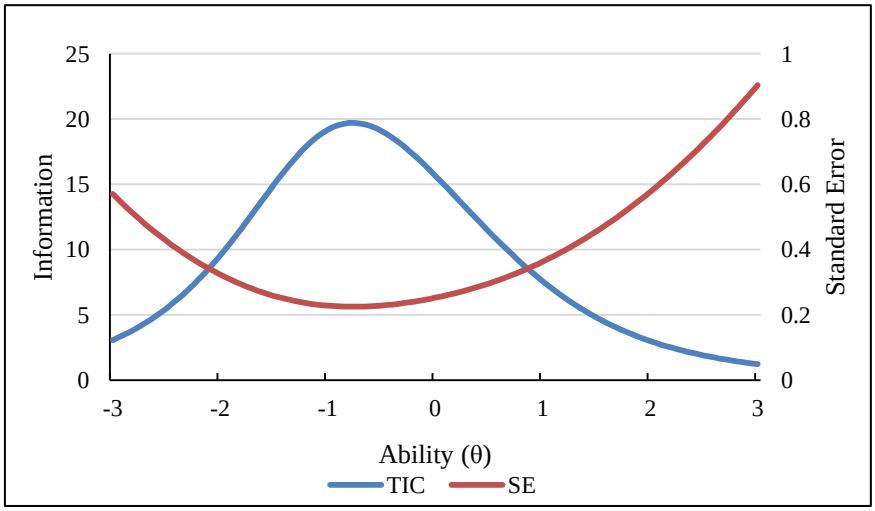

Figure 7 Test Information Curve of Form B

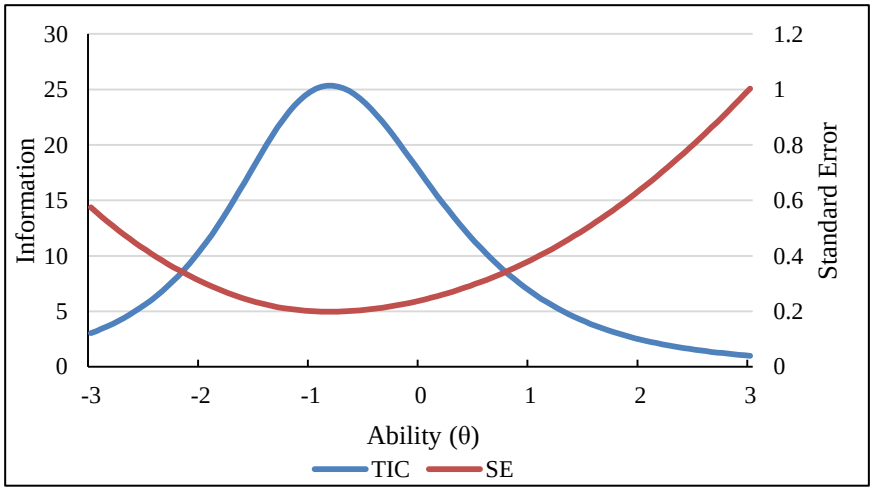

Figure 8 Test Information Curve of Form C

Figure 6 showed that Form A had smaller standard error across the ability scale from -1.9 to +0.7 and larger standard error at the low and high ends of the scale. The estimation of the student's ability was more precise across from -1.9 to +0.7 and less precise at the low and high ends of the scale. The test was best suited for the students having English ability of -0.75 because the test information was highest at that point having the value of 32.06. And then, it was observed that the empirical reliability of Form A was 0.96.

Figure 7 showed that Form B had smaller standard error across the ability scale from -2.1 to +0.9 and larger standard error at the low and high ends of the scale. The estimation of the student's ability was more precise across from -2.1 to +0.9 and less precise at the low and high ends of the scale. The test was best suited for the students having English ability of -0.7 because the test information was highest at that point having the value of 19.67. And then, it was observed that the empirical reliability of Form B was 0.94.

Figure 8 showed that Form C had smaller standard error across the ability scale from -2.1 to +0.8 and larger standard error at the low and high ends of the scale. The estimation of the student's ability was more precise across from -2.1 to +0.8 and less precise at the low and high ends of the scale. The test was best suited for the students having English ability of -0.75 because the test information was highest at that point having the value of 25.26. And then, it was observed that the empirical reliability of Form C was 0.95.

Constructing Norm Tables

In this study, norm tables were developed by transforming raw scores to scaled scores because raw scores only are not meaningful to compare the students' achievement level within a group. The scaled scores such as Percentile Rank, Stanines and z-scores and IRT true scores (τ) were used in this study.

Transformation of Raw Scores to Percentile Ranks

Table 2 Norm Table for the Three Test Forms by Percentile Ranks

Raw Scores	Percentile Rank of Form A	Percentile Rank of Form B	Percentile Rank of Form C
50	>99	>99	>99
49	99	99	99
48	97	98	97
47	96	96	95
46	94	94	93
45	91	93	90
44	90	92	87
43	87	90	83
42	84	87	80
41	81	85	78
40	78	81	76

The above table shows only a part of the transformation of raw scores to percentile ranks.

Transformation of Raw Scores to z-scores

Table 3 Norm Table for the Three Test Forms by z-scores

Raw Scores	z-scores (μ=0, SD=1)		
	Form A	Form B	Form C
50	+1.88	+2.03	+1.79
49	+1.79	+1.93	+1.69
48	+1.70	+1.83	+1.60
47	+1.61	+1.73	+1.51
46	+1.52	+1.63	+1.41
45	+1.43	+1.54	+1.32
44	+1.34	+1.44	+1.22
43	+1.24	+1.34	+1.13
42	+1.15	+1.24	+1.03
41	+1.06	+1.14	+0.94
40	+0.97	+1.05	+0.84

The above table shows only a part of the transformation of raw scores to z-scores.

Transformation of Raw Scores to Stanines

Table 4 Norm Table for the Three Test Forms by Stanines

Stanine (μ=5, SD=2)	Raw Scores		
	Form A	Form B	Form C
9	47 – 50	47 – 50	47 - 50
8	44 – 46	43 – 46	45 – 46
7	40 – 43	38 – 42	41 – 44
6	33 – 39	33 – 37	35 – 40
5	25 – 32	25 – 32	28 – 34
4	19 – 24	21 – 24	21 – 27
3	15 – 18	17 – 20	17 – 20
2	13 – 14	13 – 16	14 – 16
1	1 – 12	1 – 12	1 – 13

Transformation of IRT Ability Score (θ) to True Scores (τ)

Table 5 Norm Table for the Three Test Forms by True Scores (τ)

Raw score	Ability θ scaled score	True scores (τ)		
		Form A	Form B	Form C
49 – 50	+4.0	50	50	50
47 – 48	+3.0	47	47	48
45 – 46	+2.5	47	46	47
43 – 44	+2.0	45	45	46
40 – 42	+1.5	44	43	44
38 – 39	+1.0	41	40	42
35 – 37	+0.5	38	37	39
31 – 34	0	32	31	33
24 – 30	-0.5	25	25	27
18 – 23	-1.0	16	18	19
12 – 17	-1.5	10	12	12
9 – 11	-2.0	6	8	8
4 – 8	-2.5	3	5	5
1 – 3	-3.0	2	3	3

Construction of a Computer-Based Test with QuizFaber

In this study, computer-based test was constructed for Grade 9 English Achievement Test by using the QuizFaber Computer Software (programmed by Luca Galli). In the constructed CBT, 100 items of Grade 9 English Achievement Test for the three test forms have been included. Examples of designing questions are shown in the following figures.

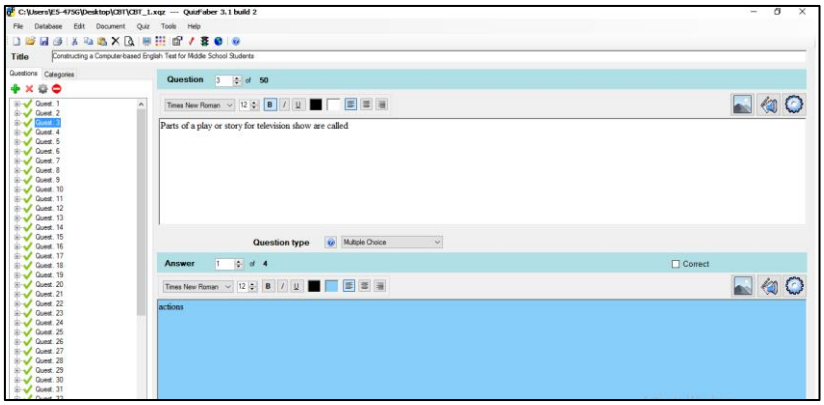

Figure 9 Question Insertion Page

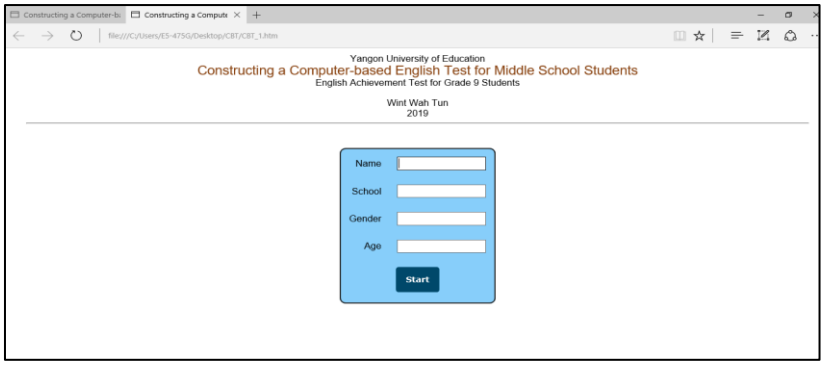

Figure 10 Student Registration Page

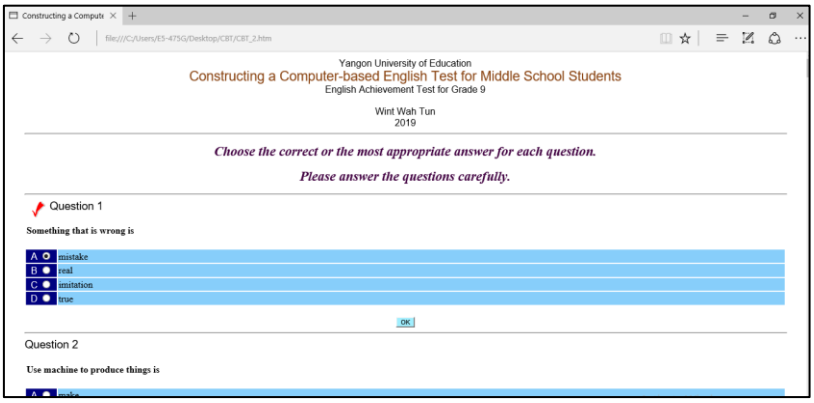

Figure 11 CBT Administration Page

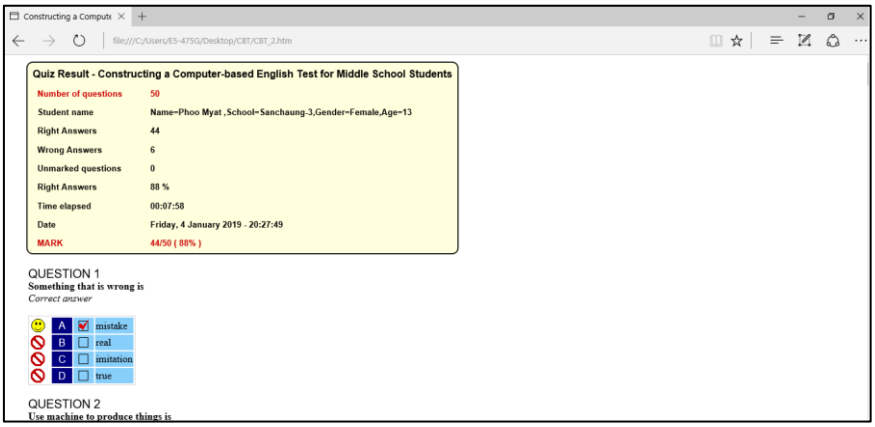

Figure 12 Final Result Page of CBT

Construction of an Online Computer-Based Test with Google Forms

In this study, Google Forms application was chosen to construct an online computer-based test (English Achievement Test for Middle School Students). QR code was used to share the test in online. This online CBT included 50 items of Grade 9 English Achievement Test. Examples of designing questions are shown in the following figures.

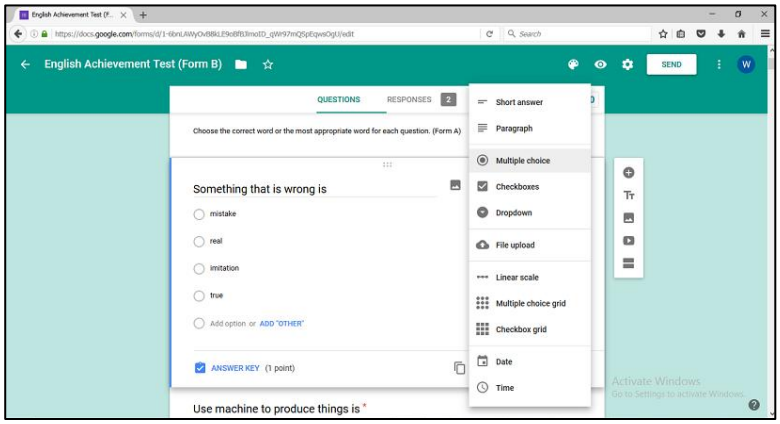

Figure 13 Question Insertion and Type of Question

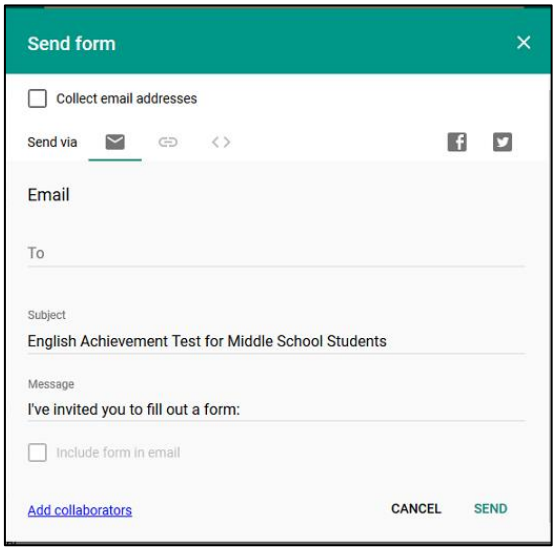

Figure 14 Distribution Page of the Test Form

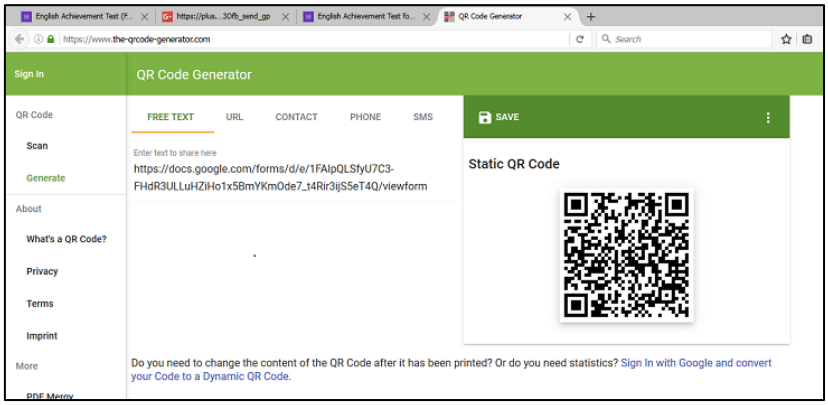

Figure 15 QR Code Generator Page

Figure 16 QR Code for English Achievement Test

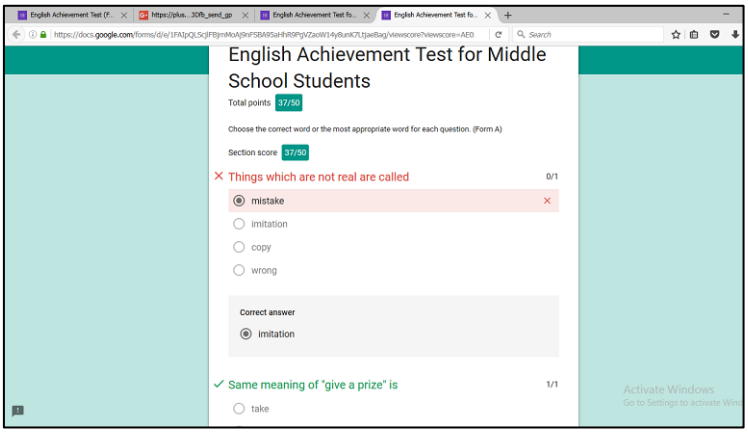

Figure 17 Student's result Page

Discussion, Further Research and Conclusion

Discussion

The main purpose of this study was to construct a Computer-Based English Test for Middle School Students. According to the results of data analysis, the main assumption of unidimensionality was firstly checked by scree plots of three forms. It was found that the three forms had reasonable unidimensionality. And for the assumption of model-data fitness, it was found that 2 PL model was fitted for the data than other models by using Lord's Chi-square method. Then, parameters of the three test forms were separately calibrated. The estimation of ability and item parameters of the three test forms were conducted by BILOG-MG 3 Software.

In the results of IRT item analysis, the mean values of *a* were 0.88 in Form A, 0.73 in Form B and 0.81 in Form C. The mean of b-values for the tests were -0.28, -0.30 and -0.45 respectively. It can be interpreted that the items of the three tests were neither too easy nor too difficult and provided appropriate discriminations.

According to the comparison of TCCs, it was found that the curves were parallel and they can be used alternatively. And, it can be said that the three test forms had appropriate difficulty and appropriate discrimination. According to the TICs, Form A was appropriate for ability (-1.9 to +0.7) of students in English. Form B was appropriate for ability (-2.1 to +0.9) and Form C was appropriate for ability (-2.1 to +0.8) of students in English. And the empirical reliability was 0.96 in Form A, 0.94 in Form B and 0.95 in Form C.

To compare the students' results, the norm tables were developed by transforming raw scores to scaled scores. The scaled scores were percentile ranks, z-scores and stanine. Moreover, the ability scores (θ) in IRT were transformed to true scores (τ) to facilitate score interpretation.

In constructing the computer-based test with QuizFaber, the items analyzed by IRT model were used and the total items were 100 multiple choice items. To create an online computer-based test, Google Forms application was used and QR code was used to send the test in online. The online computer-based test contains 50 multiple choice items.

Limitations of the Study

Some limitations were found in this study. The first limitation was sample size. Although the total sample was 1514 students, only about 500 students could be administered for each form. The sample size per test was less than 1000. So, three-parameter logistic model cannot be applied in this study. Two-parameter logistic model was used and this model can estimate the difficulty and discrimination parameters. Moreover, the tests contained multiple-choice items, so the students with low ability could choose the answer by guessing.

The second limitation was population. The population in this study was limited to Yangon Region. It is not representative of the population of Grade 9 students in Myanmar. It would be better to use other population from different regions.

Third, multiple-choice items can be constructed for computer-based test in this study. Other question format like open-ended responses are not applied in this study because they are more difficult in developing an answer key than multiple-choice questions.

Suggestions and Recommendations for Further Research

In this study, three alternate forms of Grade 9 English Achievement tests were used to construct a computer-based test. Therefore, achievement tests of other subject matters like Mathematics, Science and Myanmar should be used in the future researches to know about the tests on the other content areas. In this study, two-parameter logistic model was used and the students with low abilities can choose the correct answer by guessing. So, future studies should use three-parameter logistic model with larger samples to get more accurate results if the test contains multiple choice items.

For computer-based tests, this study involved the construction of simplest type of CBT. Constructing computer-adaptive tests is very complex, time-consuming and it needs to develop more test items. But it can give the students' test performance more accurately than the simple computer-based test. Future studies should develop other types of CBT like computer-adaptive test, multistage test and computerized classification test. It is hoped that more educational testing should use more qualified computer-based tests to increase the validity of testing and to meet the current educational trend.

Conclusion

Computer-based tests are more popular in today's world. Dozens of admissions, placement, certification, and licensure testing programs are administered on computer with the number growing each year. Computer-based tests provide ease and flexibility in administering and grading tests and allowing for the development of technology-based environment. Moreover, they not only enable the examination of objectivity, fairness, but also provide the quick results to students and teachers. Therefore, this study focus on constructing a computer-based test English

test for middle school students. It is believed that this study will provide useful information for the educators in constructing and using a computer-based test in the testing area of Myanmar.

References

Crocker, L.M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College Publishing

Demars, C. (2010). *Item Response Theory: Understanding Statistics Measurement*. Oxford University, Inc.

Google Forms Application. (2012). Retrieved January 14, 2019 from http://en.m.wikipedia.org>wiki>GoogleForms

Hambleton, R.K., & Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA:Sage.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices (2nd ed.)*. New York: Springer-Verlag.

Lin, C.-J. (2008). Comparisons between Classical Test Theory and Item Response Theory in Automated Assembly of Parallel Test Forms. *Journal of Technology, Learning and Assesement, 6* (8). Retrieved August, 4, 2018 from http:// www.jtla.org

Lord, F.M. (1980). *Applications of IRT to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Maximo. R. (1997). Norming and Norm-referenced Test Scores. *Paper presented at the Annual Meeting of the Southwest Educational Research Association*, Ausitn, TX, January, 1997.

Naw Zin Win (2017). *Constructing a Computer-based Test for Grade 9 Mathematics*. Unpublished Master Thesis. Myanmar: Yangon University of Education.

Nu Nu Khaing (2015). An empirical IRT approach to item banking and computer-based testing. *Journal of Myanmar Academy of Arts and Science*, Vol.XIII,440-447.

Olumorin, O.C. (2013). Computer-based tests; a system of assessing academic performance in University of Lorin, Nigeria. *American Academic & Scholarly Research Journal, 5*(2).

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T.C. (2001). *Practical Considerations in computer-based testing*. New York, NY: Springer.

Parshall, C.G., Harmes, J.C., Davey, T., & Pashley, P.J. (2010). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W Glas (Eds), *Elements of adaptive testing* (215 – 230). New York, NY: Springer.

Patil, F., Bhandari, U., & Kasar, M. (2015). QR Code Approach for Examination Process. *International Journal on Recent and Innovation Trends in Computing and Communication, 3*(2), 633-636.

QuizFaber computer software. (2016). Retrieved October 23, 2018, from http:// www.quizfaber.com

Simin S.,& Heidari. A. (2013).Computer-based assessment: pros and cons. *Elixir Edu. Tech. 55* (2013)12732-12734.

Wiersma, W., & Jurs. S.G. (1990). *Educational Measurement and Testing (2nd ed.)*. United States of America: Allyn and Bacon.

Zimoski, M., Muraki, E., Mislevy, R.J., & Bock, R.D. (2003). BILOG-MG 3: *Item analysis and test scoring with binary logistic models*. Chicago, IL: Scientific Software. [Computer Software]