CONSTRUCTION OF PARALLEL FORMS OF GRADE 10 CHEMISTRY AND FINDING ITS RELIABILITY COEFFICIENT OF EQUIVALENCE

Thae Phyu Htwe[1], Tun Thein[2]

Abstract

The main purpose of this study was to develop two parallel forms of test for Grade (10) Chemistry. This thesis was mainly concerned with planning, constructing, and analysis of test items. And then, two parallel forms of test were administered to 524 Grade (10) students from Ten Basic Education High Schools in Yangon Region. Descriptive survey method and quantitative research design were used in this study. From 80 items of both forms, 12 pairs of items were selected as parallel items according to their similar levels and ranges of difficulties and discriminations. Since reliability coefficient of equivalence of Form A and Form B is 0.703, both forms have strong positive relationship. So, students who performed well in Form A would also perform well in Form B. Form A can be used to get more information for measuring chemistry achievement of students who have $\theta = -1.4$. Form B had smaller standard error across the ability scale from -2.5 to +1.5 and larger standard error had at the low and high ends of the scale. The maximum amount of information was $I(\theta) = 6.7$ at $\theta = +0.2$. Ability estimates were more precise across the ability scale from -2.5 to +1.5 than at the high and low ends of the scale. Therefore, Form B could be suitable to measure for students whose ability is $\theta = +0.2$. The expected ability distributions of the students (Form A and Form B were applied) were normally distributed across the ability scale.

Keyword: Parallel Forms of Tests, Reliability, Reliability Coefficient, Coefficient of Equivalence

Introduction

There is considerable controversy about the extent of testing and about the fact that some very important decisions are based on test results. But educational tests and other measurement devices are a useful and essential part of teaching and learning. In order to get the necessary information about each student, the teachers need to select or create appropriate tests. A teacher

[1.] Tutor, Department of Educational Psychology, Yangon University of Education.
[2.] Lecturer, Department of Educational Psychology, Yangon University of Education.

uses a wide variety of test formats in order to tap students' different skills, memory, creativity, ability to organize information and the like.

Assembling equivalent test forms with minimal test overlap across forms is important in ensuring test security. Parallel-forms reliability compares two different experiments that used the same content. An increase in the number of items (tasks) to be assessed might make activities more homogenous (Dunbar, Koretz, & Hoover, 1991). Parallel forms of test can be used to find the reliability and to assess the effectiveness of instruction and training programs. Parallel forms of tests are useful to prevent knowing prematurely the test paper. Parallel forms of tests are much easier to develop for well-defined characteristics. For example, achievement tests that are given to students at the beginning and end of the school year are alternate forms. Parallel forms of tests are essential for constructing pretest and post-test to measure the effectiveness of instruction in schools and training programs. So, it is important that the teacher should acquaint with the nature, construction and uses of parallel forms of test.

Science is everywhere in today's world. It is part of our daily lives, from cooking and gardening, to recycling and comprehending the daily weather report, to reading a map and using a computer. So, children need to understand that or to be taught to think critically or provided with the tools to analyze and test a problem or situation. Chemistry is at the heart of environmental issues. Chemistry is important because everything you do is chemistry! (Helmenstine, A.M., 2014). Chemistry is sometimes called "the central science" because it bridges other natural sciences like Physics, Geology, and Biology with each other. Students wanting to become doctors, nurses, physicists, nutritionists, geologists, pharmacists, and (of course) chemists all study chemistry because chemistry related jobs are plentiful and high-paying.

All schools, including government schools, comprehensive schools, and private boarding schools have taken the University Entrance Examination, commonly referred to as the matriculation exam administered by the Myanmar Board of Examinations. Students are administered a combination of 6 tests depending on their tracks: arts, science, and arts and sciences. The subjects offered are Myanmar, English, Mathematics, Chemistry, Physics,

Biology, History, Geology, Economy and Optional Myanmar. Matriculation examination is important for students for joining universities and choosing career for their whole-lives. In order to achieve success in the matriculation examination, Grade (10) is important as a basic for matriculation examination. Therefore, this research studied how to develop Grade 10 Chemistry test and proposed about finding reliability of equivalence of the two parallel forms of tests.

Purposes of the Study

The main purpose of the study is to develop parallel forms of Grade 10 Chemistry. The specific objectives of the research are to construct two parallel forms of Grade (10) Chemistry by applying IRT technique and to find the reliability coefficient (of equivalence) of the test.

Definitions of the Key Terms

Parallel Forms of Tests: Two tests that follow the same test plan but have different content in which the items are the same in type, cover the same content, have the same distribution of difficulty values, and yield scores having the same mean, variability, and reliability. (Ebel,Robert L, 1962)

Reliability: The reliability of a test refers to the consistency with which it yields the same rank for individuals who take the test more than once. (Kubiszyn.T & Borich.G, 2007)

Reliability Coefficient: The reliability coefficient can be defined as the correlation between scores on parallel test forms. (Crocker.L & Algina.J, 1986)

Coefficient of Equivalence: The coefficient of equivalence is the correlation coefficient between the two parallel forms of test. (Crocker.L & Aligna.J, 1986)

Review of Related Literature

The Nature of Parallel Forms of Test

Suppose that all candidates for entry into a particular health occupation must take a state board examination, which is administered under controlled conditions at a particular site on a given date. To reduce the possibility of

cheating, examinees in adjacent seats take different forms of the test covering the same content. Clearly each examinee has the right to expect that his or her score would not be greatly affected by the particular form of the test taken. In this case, the errors of measurements that primarily concern test users are those due to differences in content of the test forms. Of course, administration and scoring errors, guessing, and temporary fluctuations in examinee's performance may also contribute to inconsistency of scores.

The alternate form method requires constructing both forms to the same group of examinees. The forms should be administered within a very short time period, allowing only enough time between testings. So that examinees will not be fatigued. It is considered desirable to balance the order of administration of the forms so that half the examinee are randomly assigned to form 1 followed by form 2, where as the other half take form 2 followed by form 1.

Any test that has multiple forms should have some evidence of their equivalence. Typically, tests of achievement and aptitude are constructed with multiple forms since some clinical, educational, or research uses require the examinee to have an opportunity to retake the examination, and the test user does not want to use the same items for the second test. Although there are no hard, fast rules for what constitutes a minimally acceptable value for alternate form reliability estimates. (Crocker & Algin, 1986)

The Most Important Qualities of Parallel Forms of Test

Alternate forms of a test should be thought as forms built according to the same specifications but composed of separate samples from the defined behavior domain. Thus, two parallel tests should contain questions of the same difficulty. The same sorts of questions should be asked; for example: there should be a balance of specific fact and general idea questions. The same types of passages should be represented, such as expository, argumentative, and aesthetic but the specific passage topics and questions should be different.

It is important that characteristics of a good test will be considered by test constructors. In judging the quality of a test some of the important factors-relevance, balance, efficiency, objectivity, fairness and speediness need to be

considered. Among them, validity and reliability are two main characteristics. These two are generally accepted universals.

Validity

Validity may be defined as the degree to which a test measures what it is intended to measure. The validity coefficient enables us to determine how closely the criterion performance could have been predicted from the test scores.

The American Psychological Association's Standard for Educational and Psychological Testing (1985) and Psychological testing specialists generally recognize three ways of deciding whether a test is sufficiently valid to be useful. There are content validity, criterion-related validity, and construct validity.

"Content validity involves essentially the systematic examination of the test intent to determine whether it covers a representative sample of the behavior domain to be measured."

"A test that has high content validity can be built by (i) identifying the subject matter topics and behavioral outcomes to be measured; (ii) building a table of specifications; which specifics the sample of items to be used, and (iii) Constructing a test that closely fits the table of specifications. These are the best procedures we have for measuring high content validity." As a result, it is evident that one way to ensure high content validity is to prepare a detailed specifications table or blueprint for the examination. Alternate form of a test must parallel each other in both content and difficulty. If the forms measure different content, they cannot be used interchangeably. If one form is easier than another, a passing score has different meaning for the respective forms and students will be classified as a function of test difficulty rather than degree of competence. So it is importance to develop alternate forms that parallel in content validity.

Reliability

Reliability of a test refers to the extent to which it consistently measure what is supposed to measured.

Sometimes, two or more equivalent forms of a given test will be developed to increase flexibility in administering the test. The availability of parallel forms allows retesting without worrying about practice effects. Or a teacher may wish to develop two forms of a test to reduce the likelihood of individuals copying answer from students seated nearly.

Whenever two or more forms of a test are developed with the intent of using these forms interchangeably, it is desirable to compute parallel-form reliability. This verifies the alternate test forms are measuring the something. To estimate parallel-form reliability, both forms of the test are administered to the sample people. The correlation between scores on the two forms indicates the degree to which they have parallel form reliability.

Although establishing parallel-form reliability for classroom tests is preferred when multiple forms are used, a teacher seldom has the opportunity to administer every form to each student. Therefore, alternative solutions are needed.

If the purpose of using the multiple forms is simply to control copying, a preferred orders on the two forms by using the same items but in different orders on the two forms. Probably the simplest procedure for altering the order of items is to begin a second test form with items located near the middle of the first form. Research has shown that the order in which items are presented on a test has minimal or no effect on the scores of examinees. Consequently, there is little reason to estimate parallel form reliability when the only difference between the alternate forms is the order in which the test items are presented.

When equivalent forms are developed so that students can, at a later time, be administered a retest, then distance (although similar) items should be used on the different forms of the test. In this situation, the optimal procedure is to compute the parallel-form reliability by calculating the correlation between scores on the two forms. But again, this is usually impractical.

A reasonable alternative is to take steps to ensure the equivalence of the forms. This might include writing items from carefully assigning items that measure each skill to the respective forms of the test.

Developers of commercially prepared tests sometimes prepare two equivalent forms of a given instrument. As part of their test-development process, the developers will compute the parallel-form reliability by simultaneously administering both forms of the test to a group of students. The correlation coefficient between these examinees' scores is then computed. If the alternate forms are measuring the same skills, this reliability coefficient approaches 1.00. (Oosterhof, 1990), (Thorndike, 1991)

Procedures for Developing Parallel Forms of Test

Two or more forms of an educational test are considered to be equal or equated when practically identical scores on each are made by the same individuals or by the same ability. This means that the forms of the test must be made up of test items which parallel one another closely in difficulty. In practice, such close equality of an item difficulty in alternate forms is obtained in one of three ways.

The First Procedure

This procedure involves the preparation of large number of items covering the total range of the subject matter to be tested, on the chance that there will be a sufficient number of items at each of many difficulty levels to permit of pairing items of equivalent difficulty in the alternate forms of the test. When this is done, the alternate forms of the test may be considered roughly equal in difficulty but there will be only a very general and broad equivalence of content.

The Second Procedure

This involves the preparation of parallel items on certain selected, important concepts. One item may test the identification of concept, while the other may test the identification of an additional phase of the concept or some phase of the identification of the procedure involved.

The Third Procedure

It permits the establishment of comparable forms of tests by the use of derived scores although the complexity of the statistical techniques necessary and the variety of derived scores which are used in this way make a complete presentation impracticable at this point. It may suffice here to say that the

derived scores are so established that they have constant meanings, whether or not they are obtained on the same form or from the same pupil group, and that the method establishing a 'normalized group' is basis to the procedure.

In these procedures, the items are arranged in such a manner that the two forms represent almost exactly the same difficulty as a whole, as well as almost parallel difficulty at any given point in the test. An exact equivalence of difficulty is not demanded of each pair of items as a slight difference in difficulty for the two items of one pair may be compensated by an opposite and equivalent difference in difficulty for the item of another pair. This method of shifting and balancing the items for the two forms of the test results in a roughly scaled test of two or more forms composed of items likely to be failed by approximately the same percentages of cases. The accuracy of this method of equating test forms depends to a large degree up on the extent and the representative nature of the sampling of pupil responses used in the preliminary evaluation of the items.

General Suggestions Concerning Parallel Forms of Test Construction

The general suggestions concerning parallel forms of test construction are the following.

- Prepare a preliminary draft of the test based on the table of specifications.
- Include more items in the first draft of the test than will be needed in the final form.
- After some time has elapsed, the test should be critically reviewed in order to check the items with the original outline.
- The items should be phrased so that the content rather than the form of the statement will determine the answer.
- The difficulty level of the items should be appropriate to the group of examines.
- The item content should be determined by the importance of the subject matter.
- Classroom tests should be power tests, not speed test keep tests short enough so that all students can finish.

- Keep the reading level low.
- The test may include more than one type of item.
- All the items of a particular kind should ordinary be placed together in the test.
- To the extent that it is feasible, items of a particular type should be arranged in ascending order of difficulty.
- The directions to the pupil should be as clear, complete, and concise or possible.
- Before the actual scoring begins, prepare answer keys and scoring procedures.

Methodology

Sample of the Study

This study used two parallel forms (Form A and Form B) of Chemistry achievement test for Grade (10). This study is geographically restricted to Yangon Region. Ten Basic Education High Schools were selected for this study. Participants in this study are Grade (10) students from the selected schools within the academic year (2013-2014). In each selected school, 50 students participated in this study. The content area was limited to seven chapters from Grade (10) Chemistry textbook to investigate the item qualities based on students' responses.

Instrumentation

In this study, two parallel forms of test for Grade (10) Chemistry were constructed under the direction and guidance of experts in educational test and measurement field, experts in educational methodology department and experienced teachers in chemistry department with the reference of Grade 10 Chemistry textbook and some GCE O level questions. The type of test items that is used in two forms is multiple-choice items with four alternatives.

First, seven chapters from given content of Grade 10 Chemistry Text Book were selected and multiple-choice (MC) items were constructed systematically according to rules of construction. Second, about 120 items were selected from total number of 164 items. After preparing the table of

specifications, expert review was conducted for face validity and content validity by 6 experts in the educational psychology department, department of educational methodology, and chemistry department in Yangon Institute of Education (YIOE). Next, revisions in wording and length of items were made according to supervision and editorial review of these experts. And then, each form of test was administered to 135 students from B.E.H.S (1) Hlaing in Yangon City for pilot testing. And then, some incorrect or ambiguous items and unfair items were corrected or reworded and removed based on the results of scattered diagrams.

All the items in each form were multiple-choice items. Each form consists of 40 multiple-choice items. The two forms of test were constructed in such a format that it covered the four areas based on Grade (10) Chemistry Text Book concerning knowledge, comprehension and application levels of educational outcomes.

Achievement test should measure an adequate sample of the learning outcomes and subject matter content included in the instruction. However, it is difficult to cover all the instructional objectives and all the content of course in the limited time available for testing. Therefore, a sample of items selected from the topics of instruction must be representative. One way to get greater assurance is planning a table of specifications. Both forms of test were prepared based on the same table of specifications.

Selection of Townships and Schools

At first, the townships in Yangon Region were stratified on the basis of geographical region as East, West, South and North. Schools in respective townships were selected based on the (2012-2013) matriculation pass percentage. Selected Ten Basic Education High Schools are listed in Table 3.1.

Table 3.1: Selected Schools and Number of Students from Yangon City for Administering Form A and Form B

District	Name of Schools	Pass % (Matric)	Number of Students			Total Students
			Male	Female	Total	
East	B.E.H.S (1) Thingungyun	51.13%	24	26	50	152
	B.E.H.S (2) Tamwe	46.60%	17	35	52	
	B.E.H.S (5) North Okkalapa	38.60%	17	33	50	
West	B.E.H.S (2) Kamaryut	89.17%	26	24	50	150
	B.E.H.S (3) Bahan	25.58%	25	25	50	
	B.E.H.S (4) Hlaing	41.44%	23	27	50	
South	B.E.H.S (2) Thanlynn	39.64%	29	27	56	106
	B.E.H.S (3) Thanlynn	28.57%	22	28	50	
North	B.E.H.S (1) Hlaing Thar Yar	18.68%	22	44	66	116
	B.E.H.S (3) Insein	28.83%	19	31	50	
Total	10		224	300	524	524

Test Administration

After the tests required have been formed, the next step is to prepare for the administration. Each form of test is divided into three subtests as knowledge, comprehension and application. Each form contains 40 items. Both forms of the test were administered to 524 Grade (10) students. The data so obtained was analyzed for obtaining the pairs of items based on their levels and respective areas.

Table 3.2:Showing the Difficulty and Discrimination Indices for the Knowledge Level Items in the Content Area 'Formula, Equation & Naming'

No	Subtest	Item No	P_H	P_L	Difficulty	Discrimination
1	1	2.1	0.44	0.25	0.345	0.19
2	1	2.2	0.67	0.25	0.46	0.42
3	1	2.3	1	0.5	0.75	0.5
4	1	2.4	0.84	0.5	0.67	0.34
5	1	2.5	0.72	0.42	0.57	0.3
6	1	2.6	0.31	0.06	0.185	0.25
7	1	2.7	0.89	0.36	0.625	0.53
8	1	2.8	0.64	0.56	0.6	0.08
9	1	2.9	0.44	0.39	0.415	0.05
10	1	2.10	0.83	0.31	0.57	0.52
11	1	2.11	0.47	0.25	0.36	0.22
12	1	2.12	0.78	0.14	0.46	0.64
13	1	2.13	0.78	0.36	0.57	0.42
14	1	2.14	0.11	0.11	0.11	0
15	1	2.15	0.69	0.17	0.43	0.52
16	1	2.16	0.64	0.25	0.445	0.39

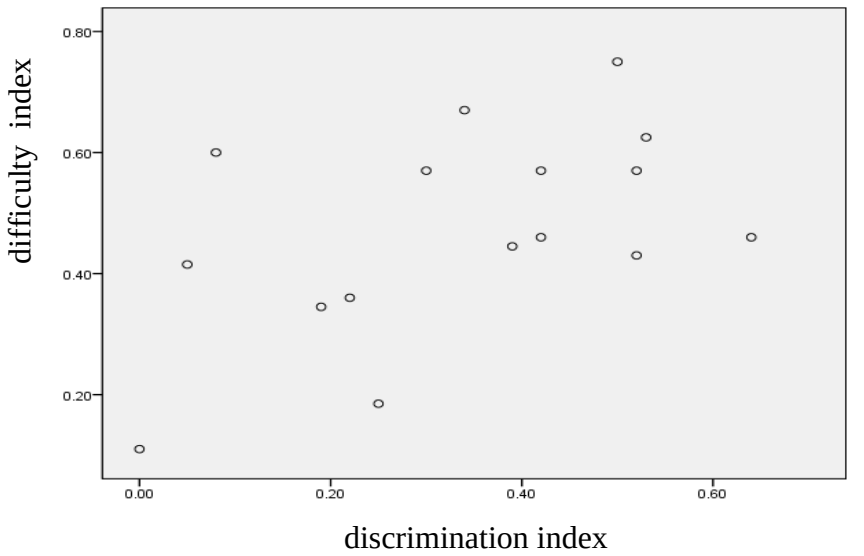

Figure 3.1: Selection of Knowledge Level Items in the Content Area 'Formula, Equation & Naming' with respect to Similarities in their Difficulty and Discrimination

Before the actual equating procedure, the descriptive statistics for Form A and Form B 2 were calculated using SPSS 16.0 programs. The distributions of scores on each form are shown in Figures 3.2 and 3.3. Descriptive statistics of both forms are described in Table 3.3.

Form A

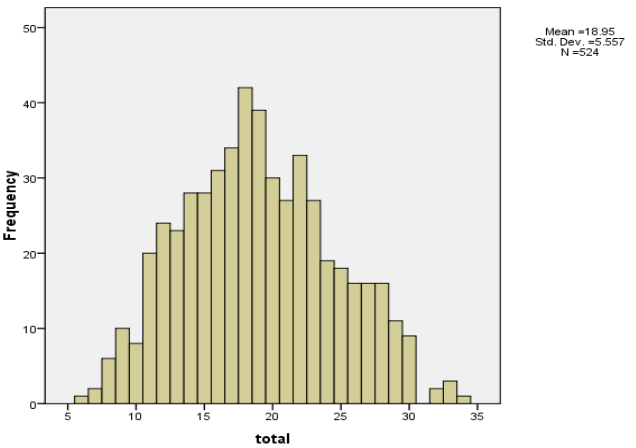

Figure 3.2 : Score Distribution of Chemistry Test Form A

Form B

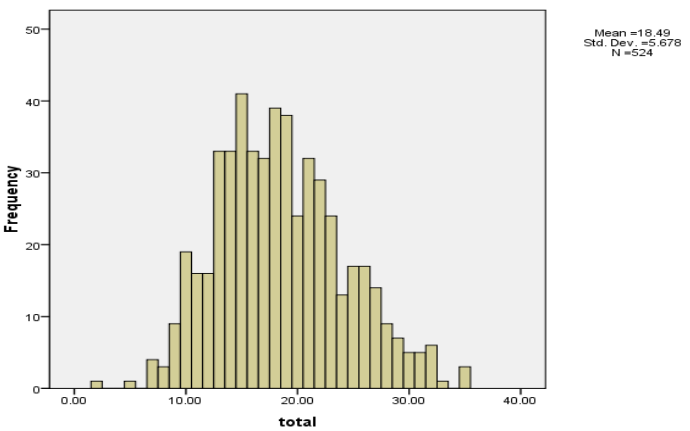

Figure 3.3: Score Distribution of Chemistry Test Form B

Table 3.3 : Comparison of Form A and Form B (Total)

	Form A	Form B
N Valid	524	524
Missing	0	0
Mean	18.95	18.491
Std. Deviation	5.557	5.678

According to Table 3.3, the means and standard deviations of Form A and Form B were little different. This fact showed that the ability level of the students were not too different attempting in both forms.

Data Analysis and Findings

Data Analysis for Checking Assumptions of Equating Test Scores

To be able to meet the assumptions of equating test scores, two forms of chemistry achievement test (Form A and Form B) were developed. The descriptive statistics of two forms of chemistry test are described in Table (4.1).

Table 4.1: Descriptive Statistics of Form A and Form B

Test Form	Number of Examinees	Scale	Mean	Std. Deviation
Form A	524	Raw	18.95	5.557
Form B	524	Raw	18.491	5.67827

Before equating Form A and Form B, the raw score means of these two forms were 18.95 and 18.491. It may be interpreted that Form A was slightly different in level and range of difficulty to Form B in measuring student's chemistry achievement even though it was tried to meet the assumptions of equivalent content and statistical specifications before actual equating procedure. As a consequence, any comparison of two test forms would be unfair for the group. Therefore, equating, the statistical method, is necessary to adjust the differences between test scores obtained from two forms due to forms difficulty. By doing so, these forms can be used interchangeably in any time of examination and the test scores of examinees took different forms can be compared.

Investigation of Phase 2 Output for Test Equating

Since two forms of chemistry test were analyzed by 2PL model in this study, so there was no c or guessing parameter for these items. The results of the item parameter estimation of both forms are described in Table 4.2.

Table 4.2 Item Parameter Estimates for Form A and Form B

Test A			Test B		
Item	***a***	***b***	**Item**	***a***	***b***
1	0.379	-0.623	1	1.78	0.347
2	0.367	3.355	2	0.501	-1.065
3	0.678	-1.014	3	0.857	-1.553
4	0.734	-0.788	4	0.46	0.926
5	0.485	-1.069	5	0.226	-0.946
6	0.463	-1.434	6	0.415	-0.703
7	0.513	-0.784	7	0.279	-0.361
8	0.584	-0.171	8	0.375	0.27
9	0.372	1.377	9	0.416	-0.002
10	0.494	-0.725	10	0.276	0.597
11	0.442	-0.264	11	0.774	-0.203
12	0.845	-1.713	12	0.889	-2.188
13	0.448	0.909	13	0.334	1.731
14	0.697	0.091	14	0.271	1.949
15	0.715	-1.233	15	0.824	-1.616
16	0.215	3.044	16	0.209	2.736
17	0.479	-0.561	17	0.452	0.17
18	0.363	-0.421	18	0.242	2.629
19	0.439	0.951	19	0.222	3.014
20	0.237	3.955	20	0.345	0.53
21	0.439	0.364	21	*	*
22	0.557	-0.217	22	0.365	0.784
23	0.182	3.523	23	*	*
24	0.486	-0.747	24	0.195	2.237
25	0.154	1.511	25	0.176	3.673
26	0.316	2.716	26	0.42	1.122

Test A			**Test B**		
Item	***a***	***b***	**Item**	***a***	***b***
27	0.137	0.854	27	0.151	1.391
28	0.203	5.133	28	0.22	3.7
29	0.381	1.693	29	0.473	-0.066
30	0.785	-1.748	30	0.64	-1.101
31	0.274	2.921	31	0.413	0.689
32	0.227	1.616	32	0.231	1.229
33	0.614	-0.3	33	0.99	-0.185
34	0.161	1.883	34	1.15	-0.349
35	0.431	-0.797	35	0.182	3.375
36	0.153	4.63	36	0.392	0.01
37	0.328	0.622	37	0.296	1.78
38	0.667	-0.675	38	0.574	-1.179
39	0.22	2.577	39	0.22	1.746
40	0.397	-0.438	40	*	*

Note. *a*: item discrimination parameter, *b*: item difficulty parameter.

good = 12
acceptable = 17
reject = 11

Table 4.3: Reliability Coefficient of Equivalence of Parallel Forms (Form A and Form B)

		Form A	Form B
Form A	Pearson Correlation	1	.703**
	Sig. (2-tailed)		.000
	N	524	524
Form B	Pearson Correlation	.703**	1
	Sig. (2-tailed)	.000	
	N	524	524

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4.4: Item Parameters for the Whole Test

Test	Parameters			
	Discrimination (*a*)		Difficulty (*b*)	
	Mean	Std.Deviation	Mean	Std.Deviation
Form A	0.427	0.190	0.700	1.839
Form B	0.473	0.336	0.649	1.555

The total information curve (TIC) gives the average probability or expected proportion of the correct as a function of the underlying latent trait. TIC is used as a replacement for the traditional concept of reliability and standard error of measurement. The standard error of the test is the inverse of the square root of information, thus, the greater information causes the smaller the standard error and the greater the reliability. Based on the results of parameter estimates of the two forms, the total information curves (TIC) for Form A and Form B were plotted.

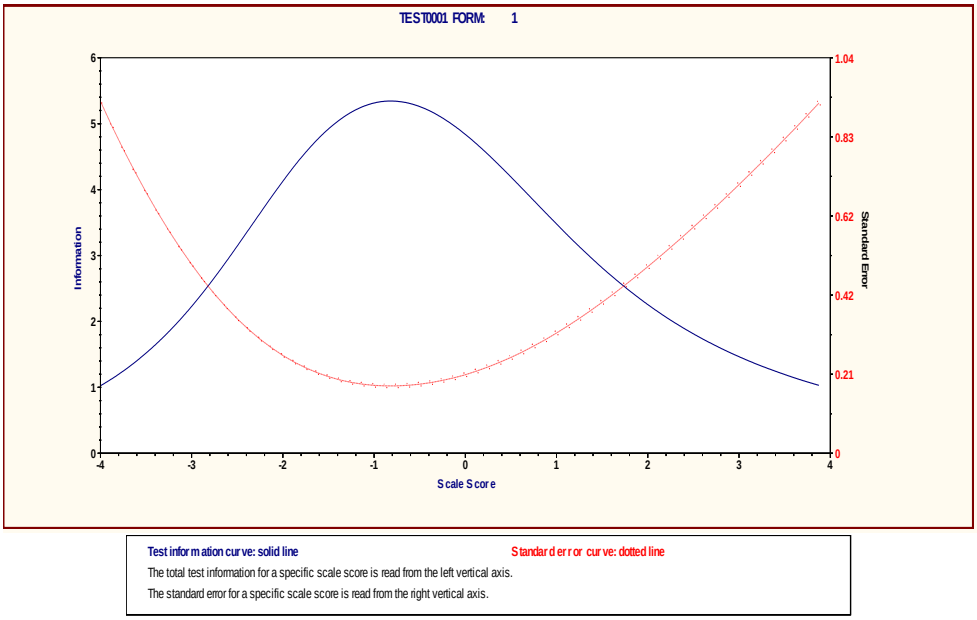

Figure 4.1: Total Information Curve of Form A

Figure 4.1 illustrated that Form A had smaller standard error across the ability scale from -2.8 to +2.7 and larger standard error at the low and high ends of the scale. The maximum amount of information was I(θ) = 6.4 at θ = -1.4. The estimation of the students' ability was more precise across from -2.8 to +2.7 than at the low and high ends of the scale. Therefore, it may be concluded that Form A can be used to get more information for measuring chemistry achievement of students who have θ = -1.4.

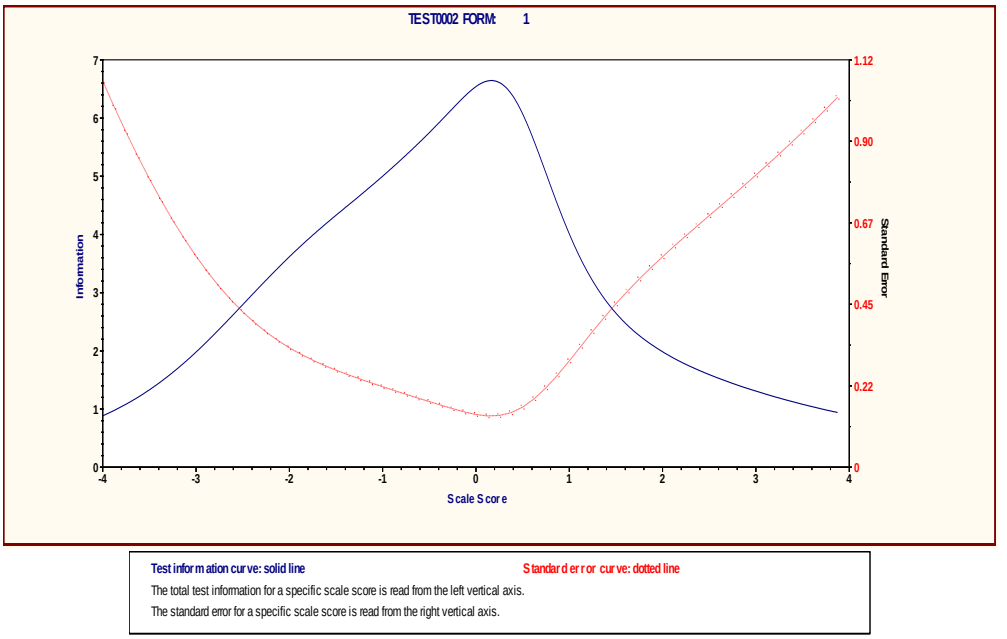

Figure 4.2: Total Information Curve of Form B

In the same way, Figure 4.2 showed that Form B had smaller standard error across the ability scale from -2.5 to +1.5 and larger standard error had at the low and high ends of the scale. The maximum amount of information was I(θ) = 6.7 at θ = +0.2. Ability estimates were more precise across the ability scale from -2.5 to +1.5 than at the high and low ends of the scale. Therefore, it may be interpreted that Form B can be suitable to measure for students whose ability is θ = +0.2.

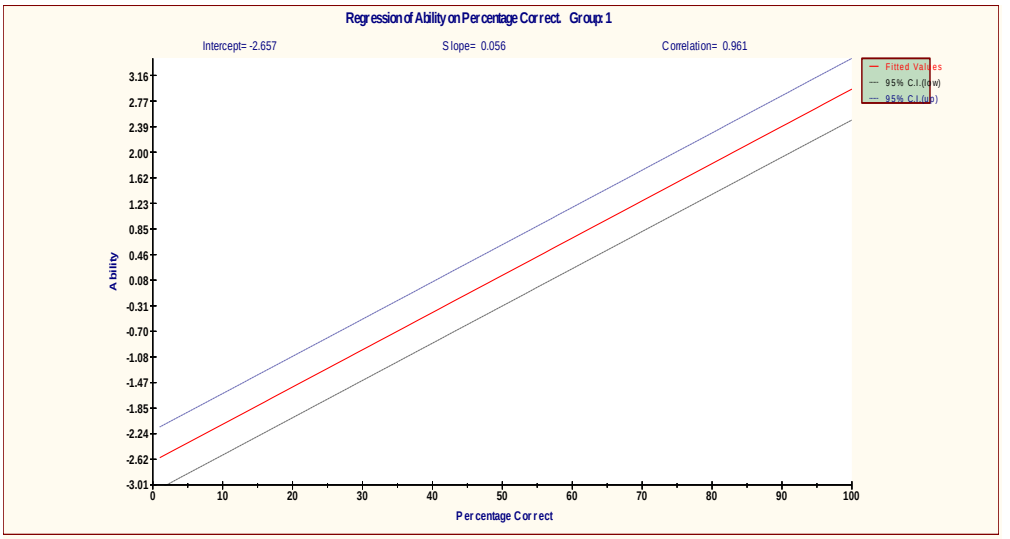

Figure 4.3: Regression of Ability on Percentage Correct for the Group

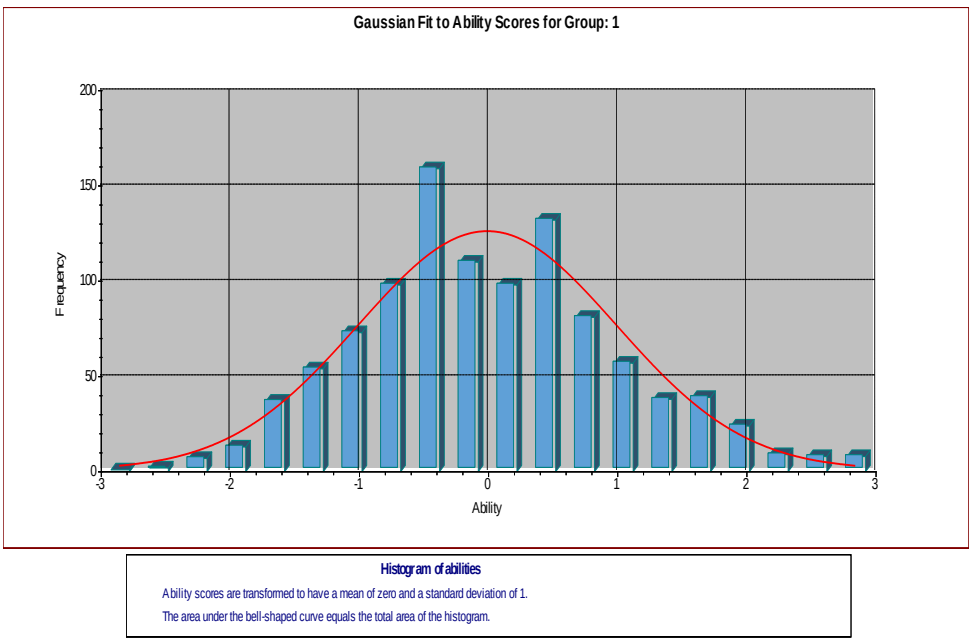

Figure 4.4 : Ability Distribution of the Students

Conclusion, Suggestion and Recommendations

Conclusion

In this study, both classical item analysis and IRT technique were used. After the test administration and scoring from pilot study, good and poor items are obtained by using item analysis technique. Then the good items which are similar not only in form and content but also in difficulty and discrimination indices were selected from four content areas to be used in the parallel forms of test. Hence, two parallel forms have been developed which have similar format, content, item difficulty and item discrimination. The total number of items constructed for the four content areas-Basic Concepts of Chemistry, Formula, equation & Naming, Solution and Gas are 120 items. From this, 80 items were selected to be used in the two parallel forms of test with the help of their respective scatter diagrams.

In this study, two forms of chemistry achievement test consisting of 40 multiple-choice (MC) items for Grade (10) students were constructed under the supervision of 6 experts in the field of education to measure the same construct. Although test forms were constructed as similar as possible to one another in content and statistical specifications, these forms might be slightly different to some extent in level and range of difficulty. Form A and Form B were administered to 524 students from selected high schools in Yangon Region. Since the sample of students were administered both forms, the researcher chose Single Group Design (Design A) of linear equating method to equate these forms. Since two forms of chemistry test were analyzed by 2PL model in this study, so there was no (c) or guessing parameter for these items. The prepared data set were entered in BILOG-MG 3 program to calibrate both forms in single run. Parallel items of both forms are selected according to their difficulty (b) and discrimination (a) values. According to Hambleton et al. (1991), the usual range for a is from 0 to 2 and high value of *a* indicates that the higher discrimination power of an item between high and low achievement of students. The values of *b* typically vary from about -3 to +3 (Hambleton,1989, Fischer & Molenaar, 1995) and the negative sign indicates that easier item difficulty and positive sign indicates that harder item difficulty. So, it can be said that these items can discriminate students who know answers from students who do not know answers. From

80 items of both forms, 12 pairs of items are selected as parallel items according to their similar levels and ranges of difficulties and discriminations. These items are item no 3, 6, 11, 12, 13, 15, 27, 30, 32, 33, 37, and 38.

Since reliability coefficient of equivalence of both forms (Form A and Form B) is 0.703, both forms have strong positive relationship. So, students who performed well in Form A would also perform well in Form B. Form A had smaller standard error across the ability scale from -2.8 to +2.7 and larger standard error at the low and high ends of the scale. The maximum amount of information was $I(\theta) = 6.4$ at $\theta = -1.4$. The estimation of the students' ability was more precise across from -2.8 to +2.7 than at the low and high ends of the scale. Therefore, it may be concluded that Form A can be used to get more information for measuring chemistry achievement of students who have $\theta = -1.4$. Form B had smaller standard error across the ability scale from -2.5 to +1.5 and larger standard error had at the low and high ends of the scale. The maximum amount of information was $I(\theta) = 6.7$ at $\theta = +0.2$. Ability estimates were more precise across the ability scale from -2.5 to +1.5 than at the high and low ends of the scale. Therefore, it may be interpreted that Form B can be suitable to measure for students whose ability is $\theta = +0.2$. The expected ability distributions of the students (Form A and Form B were applied) were normally distributed across the ability scale.

Limitations of the Study

Some limitations were found in this research. The sample selection of the students for the research was confined to Yangon Region only. Moreover, this research was performed using only the Grade (10) students and Chemistry subject. And the items used in the tests are only multiple choice items. Sample size requirement is important in test equating study. Sample size has a direct effect on random equating error. According to research of Marks and Lindsay, a small sample size is discouraged because of influencing the measure of test equating error. In this study, sample size was enough to conduct linear equating but larger sample sizes were needed for IRT equating to get more accurate equating results.

But there is no guessing parameter for multiple-choice (MC) items because the sample size of this study was less than 1000, so three parameter

(3PL) model should not be applied in this study (Lord, 1968). Since the single group design was used in this study, the performances of the examinees were affected by the order the forms are administered and practice of fatigue effects due to increased testing time. The examinees took both forms at the same time in no specific order. In order to get more parallel items, the items in tests should be constructed 1-2, 3-4,5-6 ,etc, in which 1 and 2 , 3 and 4, 5 and 6 are parallel items. By doing so, the researcher can remove poor parallel items and can select good parallel items by avoiding students' guessing and fatigue effects that can face in Form B.

Suggestions and Recommendations for Further Research

Tests play an important role in today's schools and other aspects of life. Tests poorly constructed will not give accurate information about students' achievement and hence decisions based upon this kind of information will be misleading. Teachers should use table of specifications in planning and setting classroom achievement tests. So it ensures that the test possesses content validity. Since both the essay and objective types of questions are used in almost every test, teachers should study and follow the suggestions for preparing good essay and objective questions. Teachers should develop large item pools that can be of great value when used properly.

In this study on the development of parallel forms of test for Grade (10) chemistry course, the sample is limited to Grade (10) students in Yangon Region only. It may not be a representative sample of the whole Myanmar Grade 10 student population. Thus it is necessary to conduct a large-scale research in this area on the ninth standard students of all the townships and districts in Myanmar. Future researchers should develop parallel forms of tests on science course for other grades in the primary and middle school levels which will be useful whenever a test is needed urgently. Parallel forms of test should also be developed for other subjects at different levels.

It can be pointed out that four content areas classified in this research are limited in lines with the materials covered in the selected high schools. Good items recorded in this study are quite a small number. To produce a large number of good quality items which will be of great advantage for chemistry teachers, future researchers need to develop an item bank for each topic from the ninth standard chemistry course.

Under IRT, there are many methods in test equating and different models to analyze test forms. In this study, only 2 parameter logistic (PL) model was used. So, it is recommended to apply 3PL model in test equating procedures with larger sample size to reduce equating errors. As a result of IRT equating, it was found that 12 items among 40 items are assured to be parallel. In order to get more parallel items, the items in tests should be constructed 1-2, 3-4,5-6 ,etc, in which 1 and 2 , 3 and 4, 5 and 6 are parallel items to avoid students' guessing and fatigue effects that can face in Form B.

Acknowledgements

We would like to express my deepest gratitude to the following individuals who extended their invaluable support for the completion of this study. Firstly, I would like to express my special gratitude to Dr. Aye Aye Myint (Rector, Yangon University of Education) for her precious guidance and suggestions. We would like to offer respectful gratitude to Pro-rectors, Dr. Pyone Pyone Aung and Dr. Kay Thwe Hlaing, Yangon University of Education for their administrative support that assisted greatly in the preparation of this study. We also owe a special debt to the students and the teachers from Yangon Region for their cooperation in this research.

References

Crocker,L., & Algina,J. (1986). *Introduction to classical and modern test theory*. United States of America: Holt, Rinehart and Winston, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 289-303

Ebel, R.L. (1962). Yearbook of the National Council on Measurement in Education. Retrieved January 11, 2014, from

Fischer, G. H., & Molenaar, I. W. (1995). (Eds.). *Rasch Models. Foundations, Recent Developments, and Applications.* New York: Springer–Verlag.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational Measurement. (3rd Ed.)* (pp.147–200). New York: Macmillan Publishing Company.

Helmenstine, A.M. (2014). *Why Is Chemistry Important?* Retrieved January 11, 2014, from www.chemistry.about.com/u/ua/everydaychemistry/Why-Is-Chemistry-Important.htm

Kubiszyin,T., & Borich,G. (2007). *Educational testing and measurement: Classroom application and practice* (8th ed.). United States of America: John Wiley & Sons, Inc.

Oosterhof, A.C. (1990). *Classroom applications of educational measurement*. Columbus, Ohio: Merrill Publishing Company.

Thorndike, R.L. (1991). *Measurement and evaluation in psychology and education (5th ed.).* New York: Macmillan Publishing Company.