Measuring the Reliability of the Test on Eng: 1001(2015) of Meiktila University

Lae Lae Win
Lecturer, Department of English, Meiktila University

Abstract

This study deals with measuring the reliability of the test on module Eng: 1001 of Meiktila University in 2015. The study was conducted with 54 first year non-English specialization students in the academic year 2015-16. In this study, objective test items of two different tests with the same difficulty level are taken by the same students. The scores between the odd number questions and even number questions of these students are analyzed to investigate the reliability of the test on Eng: 1001 (2015). Their scores are analyzed by using SPSS to investigate whether there is statistical difference between the scores of the test taken by the same students on different tests and between the scores of the odd numbers and the even numbers of the particular students on the target test. According to Arthur Hughes, the less different the scores , the more reliable the test. The result shows that the difference between the scores of the tests is less and the coefficient of its reliability is 0.758 and 0.874 respectively and it can be said that the test on Eng:1001of Meiktila University in the academic year 2015-16 is reliable.

Key Words_ reliability, reliability coefficient

Introduction

Teaching and testing are closely interrelated and it is impossible to carry out independence of each other. The tests can provide a clear picture on what they can do and what areas need to be improved. Tests are constructed primarily as devices to reinforce learning as a mean of accessing students' performance. Therefore, there must be clear objectives of the test – that is- which skills we mean to test and what levels we expect students be able to attain so that the test is relevant to its purpose. Therefore, teachers should aim to achieve consistent scores to provide information about the ability intended to measure, which has good impact on the instruction and learning.

The English Text Book "Straightforward" by Macmillan has been prescribed as a compulsory subject for the non-English specialization students in Myanmar since 2013. According to the test design, teaching approach changed into communicative approach and it is necessary for teachers to know how much reliable the test intended is. Therefore, this current research is carried out.

Research Questions

This research is conducted to answer the following questions,

1. Is the test Eng: 1001(2015-16) reliable or not?
2. Is there statistically significant difference between the scores of the same students on different tests as well as the scores of the odd numbers and the even numbers of the particular students on Eng: 1001(2015-16) ?

Literature Review

It is very important for teachers to design a good test because its main purpose is to measure the students' attitude towards the prescribed text , skills and knowledge in language use and language function. In order to be useful and well-qualified, the test needs to be reliable, valid and practical.

Reliability : Reliability is the consistency of measure. A measure is said to have high reliability if it produces consistent result under constant conditions. In ensuring the reliability of the test, setting clear goals and learning outcomes in teaching and assessment should be kept in mind.

Factors Affecting the Reliability of the Test

There are four factors that affect the reliability of the test; test-taker, test-scorer, test- administration and test-itself. Thus, attention to these factors can ensure the reliability of the test.

Test-scorer

This usually occurs during the subjective testing (writing and speaking) which requires judgement during the scoring process. Scoring the selective-response tests (i.e. multiple-choice) is easier than scoring productive-response ones as the answers are normally fixed and it should be sure that alternative answers are not possible.

Preparing or obtaining accurate answer key is another way of ensuring test's reliability. To ensure that the marking is objective and fair, a detailed answer key or rubrics should be provided for all markers. Rubrics are provided to teachers so they specifically know what tolook for and also to students so they know what is required of them.

Providing a detailed scoring rubric ensures the reliability of the scoring and minimizes the scorers' tedency to bias. Reliability of the scoring ensures test reliability and this is dependent test type. For example, objective testing would have perfect consistency and more reliable than subjective testing. Thus, a test should aim to permit objective scoring. Only one writing task should be given to students' ability.

Test-scorers should get enough training to score or have enough experience in scoring. Moreover, new markers should be trained to standardization exercises to be discerning in marking and clarifying acceptable responses.

Test-scorers should make some adjustments between the rubrics and the students' response because students may have some creative thinking that the scorers can't guess. At least, two markers are assigned to mark each composition scripts separately. Therefore, it is easy to compare the scores given by different markers and it helps ensure accurate assessment by checking inter-rater reliability.

Test-taker

The physical and emotional conditions of the test-takers on the exam day can affect their scores. Performances of the test-takers affect the reliability of the test. The more similar the scores are obtained, the more reliable the test is.

Test-administration

Test-administration is about the time of the day when the test is conducted, the test duration, the distracting sound and inadequate lighting etc. The invigilators should ensure that the exam room is comfortable, non-distracting and quiet, especially when a written test is conducted. Candidates should be identified with index numbers. Strict adherence to the start and end times of the test should be practised by all teachers during the test.

The number of test items should be designed appropriately to the test's duration and the students' ability. The duration of the test should not be too long or too short to minimize the test- takers' anxiety or fatigue. Sample scripts should immediately be taken after the administration of the test.

Test-Itself

One of the four factors that affect the reliability of the test is the test-itself. In this case, the teacher has to keep in mind the following ;**learning objectives, ambiguous items, clear instructions , difficulty level of the test, sufficient test items, degree of freedom.**The test needs to be based on the **objectives** as stated by the syllabus. It would also help the teachers to monitor the skills that the students should develop.

Items being set should not be **ambiguous** or biased. Questions which are biased mostly to a particular culture or gender are to be discouraged.

Students should be informed what they can expect in the test so that they know how to prepare themselves well ahead of time.

Giving **clear instructions** prevents the test-takers from spending too much time in trying to understand the instructions.

The teachers have to take into consideration for the test items to be appropriate for the **target audience** in terms of gender, race and religion.

To ensure the reliability of the test, test should not be too far from the **students' level** of knowledge and the test should not be too difficult or too easy for the students.

The test should not be highly **predicted** by the students so that teaching and learning do not concentrate only on what can be predicted.

Another thing to consider is to **provide sufficient items**, the more items are given, the more reliable the test will be.

On the other hand, teachers have to make sure that appropriate **degree of freedom** is given to the test-takers as too much freedom can make the exam unreliable.

Backwash Effect

To create beneficial backwash effect in our tests, teachers play an important role in integrating based assessment in teaching so that students can demonstrate their abilities and perform tasks at the same time. Teachers should provide opportunities that give testing to teach as opposed to teaching to test. Teachers should also include feedback for learners so that they know which areas need to improve on.
Therefore, testing as well as assessment should not be a requirement, but rather a check on the learners' progress at different stages of learning.

Reliability Coefficient

It is possible to quantify the reliability of the test in the form of a reliability coefficient. The reliability of different tests can be compared in terms of reliability coefficient. The ideal reliability coefficient is 1. A test with a reliability coefficient 1 is one which gives precisely the same results for a particular set of candidates regardless of the time administered. A test which had a reliability coefficient 0 would give sets of results quite unconnected with each other. It is between the two extremes of 1 and zero that genuine test reliability coefficients are to be found.
Lado (1961) says that good vocabulary, structure, and reading tests are usually in the .90 to .99 range, while auditory comprehension tests are more often in the .80 to .89 range. Oral production tests may be in the .70 to .79 range.

Methods and Procedures

Sampling

The participants are 54 first year non-English specialization students at Meiktila University. The cluster sampling is used to represent the whole population (all first year students of Meiktila University) and five students are randomly selected from eleven majors of non- English specialization students.

Instruments of the Study

Objective test items (Vocabulary, Grammar, Functional Language, Reading) on Eng: 1001 (2015-16) of Meiktila University and another test with the same difficulty level are used.Time alloted is 1:30 hours.

Data Collection

Five students from each specialization take two 70- item tests on one day of the two consecutive weeks. The second test is carefully designed to be the same difficulty level as the target test. The tests are carefully administered as exactly as the actual test with the support of my colleagues. The scores of the two different tests by the same students are compared by using parallel-form method in table (1), (2), (3) and the scores between the odd number questions and the even number questions of the students on the target test are compared by split-half method in table (4), (5), (6).

Research Design and Variables

This research is quantitative research and descriptive. Two independent variables , scores on the tests, are included.

Types of Reliability Used

Parallel – Forms method

The alternate test forms that are equivalent in terms of content, response processes are administered to the same group of people. Two sets of scores by the same students on different tests are analyzed by Pearson product-moment correlation coefficient of SPSS.

Split- half Method

A target test is administered to a group of individual students and split the test itself in half by even numbers and odd numbers. Spearman's Correlation of SPSS is run to determine the relationship between the two sets of scores of the odd number questions and even number questions of the target test by the same group of students.

Table 1. Comparison of the scores by the same students on different tests

Sr. No	Students	Scores on the target test 1	Scores on the test 2
1.	M1	36	32
2.	M2	53	43
3.	M3	28	44
4.	M4	32	33
5.	M5	30	37
6.	G1	48	46
7.	G2	46	42
8.	G3	39	31
9.	G4	33	30
10.	G5	31	33
11.	H1	37	32
12.	H2	50	35
13.	H3	43	40
14.	H4	30	27
15.	Os1	40	39
16.	Os2	24	30
17.	Os3	36	35
18.	Os4	22	27
19.	Os5	28	32
20.	Philo1	27	26
21.	Philo2	27	33
22.	Hhilo3	35	33
23.	Philo4	13	30
24.	Philo5	37	20
25.	Chem1	30	31
26.	Chem2	39	44
27.	Chem3	45	44

Table (1) shows the comparison of the scores by the same students on different tests to analyze by means of Parallael- forms methods.

Table 2: Comparison of the scores by the same students on different tests

Sr. No	Students	Scores on the target test 1	Scores on the test 2
28.	Chem4	52	44
29.	Chem5	63	59
30.	Phys1	54	54
31.	Phys2	58	38
32.	Phys3	64	53
33.	Phys4	59	51
34.	Phys5	56	43
35.	Maths 1	49	35
36.	Maths 2	48	39
37.	Maths 3	26	37
38.	Maths 4	31	33
39.	Maths 5	22	24
40.	Zool 1	49	42
41.	Zool 2	46	51
42.	Zool 3	51	50
43.	Zool 4	55	38
44.	Zool 5	44	40
45.	Bot 1	23	26
46.	Bot 2	52	56
47.	Bot 3	37	30
48.	Bot 4	30	37
49.	Bot 5	62	47
50.	Geol 1	44	50
51.	Geol 2	30	29
52.	Geol 3	28	28
53.	Geol 4	37	32
54.	Geol 5	31	37

Table (2) shows the comparison of the scores by the same students on different tests to analyze by means of Parallael- forms methods.

Table.3 Correlation between the scores of the same students on different tests

Test1	Pearson Correlation	1	.758**
	Sig. (2-tailed)		.000
	N	54	54
Test2	Pearson Correlation	.758**	1
	Sig. (2-tailed)	.000	

e (3) shows the correl

Table (3) shows the correlation between the scores of the same students on different tests by parallel- forms method.

Table.4 Comparison of the scores between the odd number questions and even number questions

Sr. No	Students	Scores of Odd No	Scores of Even No
1.	M1	19	13
2.	M2	26	25
3.	M3	12	14
4.	M4	18	14
5.	M5	13	13
6.	G1	17	14
7.	G2	19	18
8.	G3	20	21
9.	G4	25	20
10.	G5	17	14
11.	H1	15	14
12.	H2	20	22
13.	H3	20	28
14.	H4	16	18
15.	Os1	17	21
16.	Os2	12	8
17.	Os3	13	10
18.	Os4	13	13
19.	Os5	14	19
20.	Philo1	13	9
21.	Philo2	7	5
22.	Hhilo3	16	16
23.	Philo4	14	16
24.	Philo5	10	14
25.	Chem1	12	16
26.	Chem2	25	24
27.	Chem3	19	24

Table (4) represents comparison of the sources between the old number questions and even number questions taken by non- English specialization students.

Table.5 Comparison of the scores between the odd number questions and even number questions

Sr.No	Students	Scores of Odd No	Scores of Even No
28.	Chem4	30	29
29.	Chem5	19	16
30.	Phys1	26	25
31.	Phys2	27	27
32.	Phys3	28	29
33.	Phys4	27	32
34.	Phys5	27	30
35.	Maths 1	21	25
36.	Maths 2	11	11
37.	Maths 3	15	16
38.	Maths 4	13	13
39.	Maths 5	21	28
40.	Zool 1	23	24
41.	Zool 2	20	19
42.	Zool 3	26	28
43.	Zool 4	25	23
44.	Zool 4	25	23
45.	Bot 1	30	28
46.	Bot 2	10	11
47.	Bot 3	24	23
48.	Bot 4	12	16
49.	Bot 5	16	17
50.	Geol 1	20	20
51.	Geol 2	12	16
52.	Geol 3	18	17
53.	Geol 4	13	13
54.	Geol 5	12	15

Table (5) represents comparison of the sources between the old number questions and even number questions taken by non- English specialization students

Table.6: Correlation by Split-half Method

			group1	group2
Spearman's rho	group1	Correlation Coefficient	1.000	.874**
		Sig. (2-tailed)	.	.000
		N	54	54
	group2	Correlation Coefficient	.874**	1.000
		Sig. (2-tailed)	.000	.
		N	54	54

Table (6) shows the correlation between the scores of old number questions and even number questions by the non-English specialization students on the target test.

Findings and Discussion

Some scholars assume that the students can't be invigilated as exactly as in the authentic test. However, to get the real scores of the authentic test, it is conducted as exactly as possible. Therefore, the four factors that affect the reliability of the test: test-taker, test-scorer, test-dministration and the test itself are paid attention. They make the students take the test like the authentic test and the students want to know their real ability of how many lessons they have prepared for the coming test.

On the surface of the title of current research , it is very appealing to do . However, the time to be conducted and the participants of the research to conduct the test are very challenging.The reserch can be tackled only after the students have finished studying all the lessons and are ready to take the test. So, the test is conducted between the interval of the real test and the students' private study.The students from all specilizations of Meiktila University are to take the test to represent the whole population of Meiktila University. Therefore, the same students are persuaded to take the two tests on the consecutive weeks with the help of my colleagues so that the different sets of scores by the same students can be compared.

By means of Pearson Correlation SPSS, the reliability coefficient by parallel-form method is r = 0.758 based on n=54 observations with pairwise non-missing values. The target test and its alternative forms have a statistically significant linear relationship ($p < .001$). The direction of the relationship is positive (i.e. two tests are positively correlated) , meaning that these variables tend to increase together (i.e. greater scores on target test are associated with greater scores on alternative test).

Spearman's Correlation of SPSS is run to determine the relationship between the two sets of scores of the odd number questions and even number questions of the target test by the same group of students. There is a strong, positive correlation between the scores of odd number questions and even number questions, which is statisticaly significant (r = 0. 874, p = 0.000).

Conclusion

The reliability coefficient value of a good reading, vocabulary and structure test is 0.80-0.89 and 0.70-0.79 is adequate. It can be said that it is a good test as its reliability coefficient is 0.758 by parallel- form methods and 0.874 by split-half method. Although a test should not be rejected or selected based solely on the size of its reliability coefficient, it is a must for language teachers to determine if its reliability is acceptable and to report the reliability estimates that are relevant for a particular test.

Acknowledgements

I would like to express my gratitude to Dr. Ba Han ,Rector, Meiktila University, Dr. Tin Htun Aung ,Prorector, Meiktila University and Dr. Khin San Yu, Professor and Head of English Department for their permission, valuable advice and support to submit this research entitled " Measuring the Reliability of the Test Eng:1001(2015)".

The ones to whom I owe innumerable thanks are Dr. Ko Ko Kyaw Soe (former Rector of Meiktila University) who arranged to give the training courses of (SPSS), Research Methodology and The Art of Teaching , Associate Professor, Dr. Lay Yoon Fah (Faculty of Psychology and Education, University of Malaysia Sabah) for teaching us using (SPSS) and Research Methodology and Dr. Zeya Oo (Visiting Professor, Meiktila University) for teaching us Research Methodology and The art of Teaching. I could not have accomplished this paper without the steadfast effort, cooperation and active participation of the students of Meiktila University and my colleagues who helped me to invigilate the students and score the test.

References

Carr,N.T.(2011).Designing and Analyzing Language Tests. Oxford.

Hughes, A. (2003).*Testing for Language Teachers* (2nd Edition). New Deli: Cambridge University Press. PB71.5H841

Internet Resources

https:// www.teachingenglish.org.uk
https://www.pearsonelt.com
https://grammargraph.wordpress.com
https://hr-guide.com>data
http://languagetesting.info/
www.englishforum.com
Measuring Learning & Performance: A Primer | Retrieved from Charles DennisHale.orghttps://en.m.wikipedia.org