THE VALIDITY AND REALIBILITY OF IELTS SPEAKING RUBRIC FOR INLINGUA INTERNATIONAL TEACHERS

Purwanti Taman
Universitas Pamulang, Banten
amytaman2@yahoo.com

Abstract

The field of English for specific purposes (ESP) has developed rapidly to become a major force in English language teaching. IELTS (International English Language Testing System), as one of the ESP branch, is one of the biggest ESL tests needed world-wide.This study was conducted to find validity and reliability of IELTS speaking rubric between three teachers who assess speaking ability for the same students in the same class. This would enable IELTS teachers to measure the students' speaking ability likely in the same band area. The participants of this quantitative study were 21 Indonesian graduate students who were taking English for Pre-departure program of their Master's Degree abroad, an English for Academic Purpose (EAP) class. The result shows that the teachers in Inlingua Internatinal English training produce reliable and valid results of the speaking IELTS subtest; even though they have different working span experience because they use the same rubric. Also, the correlation between one group of scores to others are significant and linear. This can be reference for the future in the way Inlingua International maintain their teachers' marking scheme.

Keywords: *IELTS, validity, realibility, speaking rubric, ESP*

INTRODUCTION

One of the sub-skills that IELTS measures in the test is speaking ability, apart from the other three skills, listening, reading, and writing subtest. In speaking tests, the scoring of speaking performances involves the construction and application of rating scales, where each IELTS teachers have to have tacit agreement that certain speaking ability of the students should be in certain band level. This will enable the teachers to measure the students' speaking ability likely in the same band area. In other words, three teachers who assess speaking ability for the same number of students in the same class are expected to have likely the same result. Those are the reasons why the research of the speaking subtest is conducted.

Based on the speaking assessment criteria, there are four particular skills being assessed during the test namely, fluencly and easy flow, vocabulary range, grammar usage and accuracy, and pronunciation. When viewing those four influential rating scales for example, IELTS teacher have to be able to reflect the students' ability in order to know their certain band score area and to be able to improve that ability for the real test. Some teachers may award relatively lower score of their ability, while others give the score likely on their ability. This corcern comes from certain consideration that teachers should not give a false hope to the students so that they will not be disappointed if they get low score on the real test. Another consideration however, giving as minimum as could be for the band score of the students might bring demotivating effect toward them. Therefore, some teachers offer likely bandscore that belong to their ability. Indeed, this also does not mean risk free as the minimum performance migt happen in the real test will possibly cause them getting lower band score than the ones usually in preparation practice.

This issue interest the writers to evaluate further and to find out the realibity and validity of the latest IELTS speaking rubric which is used as the primary reference for teacher in measuring the speaking ability of the students. The case is evaluated based on the result of the three IELTS teachers giving speaking assessment towards 21 students in the same class. The realibility and validity of the result is measured using SPSS programme in order to find out the result for future reference for teacher assessing future students.

The purposes of people taking IELTS test are mainly known as measuring their language ability to be submitted as one of the prerequisite documents for studying or working in English speaking countries. IELTS has nine-point scoring system to define the language skill of the test-takers. The test-takers receive scores for each language skill (listening, reading, writing and speaking) and an Overall Band Score on a Band Scale from one to nine.

Speaking
The are four areas that the examiners pay attention when the test-takers taking speaking test namely: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy and Pronunciation. Those four areas are considered equally in contributing the sub-score. The remaining three other sub-scores eventually contribute the whole band score received by the test-takers within, normally, two-week time after the test date.

Nine-point scoring system

The test takers for both practice test and the real IELT test are awarded nine-point scoring system for the four skills. The nine bands reflect the participants' ability of English language which can be the reference for academic institution such universities and colleges, as well as for companies around the world. The descriptive statements of the nine-point scoring system are as follows (http://www.ieltsessentials.com):

9 Expert User – This means that the candidates has great understanding of the language, including how to use the it appropriately, accurately, and fluently.

8 Very Good – This means that the candidates – overall - use the language features very well, even though they sometimes use it inaccurately and inappropriately. Their production may cause minsunderstanding in certain unfamiliar situations.

7 Good User – This means that the candidates use the language features well, even though they produce inaccuracies, inappropriacies and misunderstandings more frequently in certain unexpexted topic discussion compared to band 8 achievers. In general, they are able to cope with complex language well.

6 Competent User – This means that the candidates use the language features effectively, in spite of some usage that is inappropriate, inaccurate, and causes misunderstanding. Overall, the users can use and understand the language failry, especially in fairly familiar situations.

5 Modest User – This means that the candidates use the language partially. They mainly be able to deliver the meaning, even though many incorrect features are made. In general, the test-takers are able to use the basic features of language in familiar situation.

4 Limited User – This means that the candidates use the language that is limited to familiar situations. They often made mistake in their production which impede the understanding. They cannot yet use complicated language.

3 Extremely Limited User – The candidates who receive this score are those who are able to deliver the message only in very familiar situations. They often make a chunk of message.

2 Intermittent User – This score is awarded to those who actually do not make real communication as they are only able to memorized based information. They mainly have great difficulty understanding both spoken and written English.

1 Non user – This score may not be awarded unless to those who do not actually have the ability to use the language except a few isolated words.

2 Did not come to the test – There is no assessment information recorded.

IELTS Speaking Subtest

Speaking sub test is conducted as soon as the test takers finish the three earlier sub tests. Those are listening which takes for 30 minutes, reading for one hour, and writing task one and two which lasted for one hour. The speaking sub test is the last test of the four which should be taken by test takers. There are four skills should be asessed in speaking sub-test for IELTS (UCLES 2009) as follow:

a. Fluency and Coherence
This measures the candidate's ability to talk with a regular pace and levels of flow, speech rate and effort, and to connect ideas and language features together in one wholistic formation.

b. Lexical Resource
This identifies how broad of vocabulary the candidate can explore and how clearly meanings and attitudes can be delivered which counts the variation of words attempted and the skill to join a vocabulary gap by pharaprasing the idea.

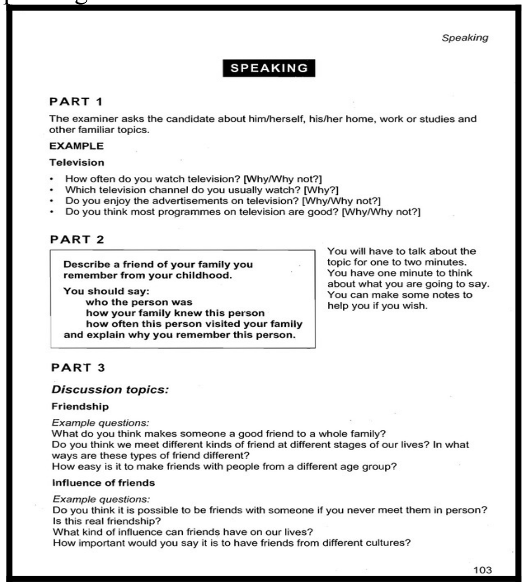
c. Grammatical Range and Accuracy
This tests the ability of the candidates in using grammar and structures accurately and appropriately. They are assessed how to produce a speech at a certain length, and level of complexity they are using.

d. Pronunciation
This refers to the candidate's ability to produce understandable speech and to apply variety of of pronunciation features to deliver meaning. The test identifies whether or not the pronunciation impedes the understanding of the listener, and also the transformation influenced by their first language.

There are three sub parts for the speaking test with variety levels of difficulty that include the skill to express opinions and information on daily conversation and middle-of-the-road experiences and situations by responding a range of questions. The ability to talk at certain length on an instructed ideas using proper language and organising thought coherently, and also the skill to share and justify views and to analyse, discuss and speculate about issues will also be evaluated in this speaking test. The test is lasted for 11-14 minutes with the following explanation for each part (IELTS for Candidate, 2009).

a. The first part is introduction and interview which lasted for 4-5 minutes. The test started by self-introduction by the examiner, followed by some confirmation question about the test-taker personal information. The process then continued by questioning the candidates about familiar topics such as hobbies, family, friends and home. questions
b. The second part of the test is an opportunity for the test-taker to talk for two minutes from the prompt card given by the examiner. They should talk about particular topic given in the card with come points they need to cover. The test-taker is given one minute to prepare their talk, and they can make a note if the they wish. The candidate talks for 1-2 minutes on the topic. The examiner usually provides one or two questions after their talk. This part lasted 3-4 minutes.
c. The third part is time when the examiner can interrupt the test-taker's talk which is called two-way discussion lasting for 4-5 minutes. Usually the examiner asks follow-up questions associated with the ideas given in the second part. These questions are considered to be more abstract that the candidate need to prepare in expressing rather complicated ideas.

The following is the example of the speaking test questions adopted from Cambridge IELTS 8 (103) Speaking Test 4:



Speaking

SPEAKING

PART 1

The examiner asks the candidate about him/herself, his/her home, work or studies and other familiar topics.

EXAMPLE

Television

- How often do you watch television? [Why/Why not?]
- Which television channel do you usually watch? [Why?]
- Do you enjoy the advertisements on television? [Why/Why not?]
- Do you think most programmes on television are good? [Why/Why not?]

PART 2

Describe a friend of your family you remember from your childhood.

You should say:
who the person was
how your family knew this person
how often this person visited your family
and explain why you remember this person.

You will have to talk about the topic for one to two minutes. You have one minute to think about what you are going to say. You can make some notes to help you if you wish.

PART 3

Discussion topics:

Friendship

Example questions:
What do you think makes someone a good friend to a whole family?
Do you think we meet different kinds of friend at different stages of our lives? In what ways are these types of friend different?
How easy is it to make friends with people from a different age group?

Influence of friends

Example questions:
Do you think it is possible to be friends with someone if you never meet them in person? Is this real friendship?
What kind of influence can friends have on our lives?
How important would you say it is to have friends from different cultures?

103

Ch

Figure 1: Speaking Question Sample

Rubrics become one of alternative to overcome difficulties in testing speaking skill which have many criteria to be scored. All types of rubrics can be used, however, analytic rubrics are considered as the best one. Analytic rubrics provide detailed information and enable to cover all features in speaking such as vocabulary, grammar, semantics (informational and fluency), and phonology (pronunciation, stress and intonation) which are all included in the four speaking assessement criteria mentioned above. The following are the characteristic of analytic rubrics:

1. The rubric objective is to clearly and distinctly identify the terminal behavior expected from a learner as the result of a learning experience.
2. The rubrics stated the performance criteria.
3. Rubrics concern the most on competence-the capacity to do what supposed to be done.
4. Rubrics can be a complete concept, contain several sub-ordiantes, or limited to specific scope as part of holistic one.
5. Rubrics should be developed together with a taxonomical framework such as found in Bloom's taxonomy..
6. Rubrics developed should concern variety of intelligences in cognitive, affective and psychomotor domains.
7. Rubrics should be developed based on learning experience, and be informed to the students before the learning activity begun.
8. The students should be involved in the development of rubrics along with their competence and maturity.
9. Instruments applied for the evaluation should provide an opportunity for the students to grow and develop learning experience.
10. The terms stated in the rubrics should be defined clearly and particularly connected to al level of cognitive ability that is constantly applied from one rubric to another.
11. Rubrics can be applied to almost all levels and any curriculum area.

Speaking Marking Scheme

Teachers teaching IELTS, like IELTS examiners, should understand the marking requlation and are requested to show their standard marking ability before conducting a real marking job and are permitted to mark all four of the IELTS subtests. This paper only discusses particularly for marking speaking. Teachers who teach speaking, ideally should be familiar with the speaking marking scheme, particularly with speaking band scores. Most of institutions normally train their teachers before allowing them to teach or even to assess and mark speaking.

Candidates receive scores from 1 to 9 of Band Scale. As an overall score from each skill. The four scores obtained from each skill are averaged and rounded to produce an Overall one.. Overall Band Scores and scores for each sub-test (Listening, Reading, Writing and Speaking) are reported in whole bands or half bands. Every score from the sub-test is equally weighted. The Overall Band Score is obtained from the mean of the total of the four individual sub-test scores. The score is then to the nearest whole or half band. The formula of rounding is provided to ensure the right conversion: if the average across the four skills ends in .25, it is rounded up to the next half band, and if it ends in .75, it is rounded up to the next whole band. Thus, a test-taker who is awarded 6.5 for Listening, 6.5 for Reading, 5.0 for Writing and 7.0 for Speaking would be awarded an Overall Band Score of 6.5 (25 ÷ 4 = 6.25 = Band 6.5). The way Overall Band Scores are marked is perfectly applicable for the ways speaking band scores are given. Based on the writer's experience, it is easier to mark especially speaking and writing by classifying the ranging scores into the band scores as follow:

Speaking Band Scores	Ranging Average Score
9	8.75 – 9.00
8.5	8.25 – 8.74
8	7.75 – 8.24
7.5	7.25 – 7.74
7	6.75 – 7.24
6.5	6.25 – 6.74
6	5.75 – 6.24
5.5	5.25 – 5.74
5	4.75 – 5.24
4.5	4.25 – 4.74
4	3.75 – 4.24
3.5	3.25 – 3.74
3	2.75 – 3.24
2.5	2.25 – 2.74
2	1.75 – 2.24
1.5	1.25 – 1.74
1	0.75 – 1.24
0.5	0.25 – 0.74

Table 1- Speaking Band Scores Coversion

METHOD

The method applied for this study is quantitative which uses SPSS program to examine the validity and the realibility of the speaking rubric. The data – speaking scores of the same students from three different teachers – gathered during the training program are then classified and inputted into SPSS software which also involved corelation test.

The Normal Distribution

Normal distribution can be difined as a bell-shaped symmetrical frequency distribution curve according to Bussiness Dictionary (2013), where two or more variables have direct relationship and high predictability or low variation. In other words, in normal distribution, extremely-large values and extremely-small values are rare and occur near the tail ends. Most-frequent values are clustered around the mean, which here is same as the median and mode, and fall off smoothly in either side of it. In normal distribution, 68 percent of all values lie within one standard deviation, 95.45 percent within two standard deviations, and 99.8 within three standard deviations. This means that there is only one out of a thousand values will possibly fall outside of six sigma. Normal distribution means the distribution has ideal or standard sense against which other distributions may be compared.

Validity and Reliabilty

The validity and reliabilty of the Speaking Band Descriptor is measured by evaluating the results of the three speaking assessments which are conducted by three different teachers. Reliability according to Phelan and Wren (2005-2006) is "the degree to which an assessment tool produces stable and consistent results". In other words, the rubric is reliable when the results of the three assessments between those three teachers are firmly fixed and are unlikely to move and change from from one to other respondence accordingly. In addition, the results also typically in agreement with other facts, in this case, with the other two assessment results. Validity, furthermore, also mentioned by Phelan and Wren (2005-2006) refers to "how well a test measures what it is purported to measure." It is important because instead of reliable, tests also need to be valid.

Correlation

Borrowing the term from Farlex (2013), "correlatinal analysis is the use of statistical correlation to evaluate the strength of the relations between variables." It can be explained further that correlation is a statistical technique that enable to identify whether and how strongly pairs of variables are related. The major result of a correlation is named the correlation coefficient ("r") which ranges from -1.0 to +1.0 (http://www.surveysystem.com/correlation.htm). The closer r is to +1 or -1, the more closely the two variables are related. If r is close to 0, this means that there is no relationship between the variables. If r is positive, it shows that as one variable gets larger the other gets larger. If r is negative it informs that as one gets larger, the other gets smaller (often called an "inverse" correlation).

While correlation coefficients are normally reported as r = (a value between -1 and +1), squaring them makes then easier to understand. The square of the coefficient (or r square) is equal to the percent of the variation in one variable that is related to the variation in the other. After squaring r, ignore the decimal point. An r of .5 means 25% of the variation is related (.5 squared =.25). An r value of .7 means 49% of the variance is related (.7 squared = .49).

Measurement and Evaluation Planning

The indicator use to evaluate students in IELTS Speaking test should consist of four areas namely Fluency and Coherence, Lexical Resources, Grammatical Range and Accuracy, and Pronunciation. Those four areas are taken from speaking band describtor which is issued by IELTS institution. Teachers measuring the speaking ability of the students use this reference to position them in certain band score area. The speaking band descriptor is attached in the end of this file.

The data is collected and calculated in aiming to find out the the results from the three times assessments are valid and reliable. Knowing the validity and the reliability of the data is beneficial for future reference. The data for this research illustrates that there are three examiners evaluating students on their level of English speaking performance in accordance to the nine band descriptions. Data shown on table 3.1 is the result on IELTS speaking test conducted by three different teachers to 21 students as population, who are doing training in one of the well-known English institution in Indonesia. The first teacher is a local teacher who has worked for the institution for more than three years, the second teacher is a native speaker who has worked for more than five years, and the last one is a local teacher who has worked for the institution for a year only. The band scores from the three teachers in the table below are adjusted overall scores which is explained in chapter 2.

- Table 2 - Students' Speaking Band Scores

No	Student Name	1	2	3
		Adjusted Overall	**Adjusted Overall**	**Adjusted Overall**
1	Student 1	5.5	5	6
2	Student 2	6	6	6.5
3	Student 3	7	7.5	7.5
4	Student 4	7	7.5	7
5	Student 5	6.5	5.5	6.5
6	Student 6	6.5	6	6.5
7	Student 7	6.5	6.5	7
8	Student 8	6.5	6.5	7
9	Student 9	6	6	6.5
10	Student 10	5	5.5	6
11	Student 11	6	6	6.5
12	Student 12	6.5	5.5	7
13	Student 13	6.5	7.5	7
14	Student 14	7	8	7.5
15	Student 15	6.5	6.5	7.5
16	Student 16	6	7	7
17	Student 17	6	6.5	7
18	Student 18	6	7	7
19	Student 19	5	6	6.5
20	Student 19	6.5	7	7.5
21	Student 20	6.5	6.5	6

DISCUSSION

Data Analysis

This section analyzes the data which particularly emphasizes on validity and reliablity of it. The process starts from evaluating that distribution of the data is normal. It then followed by the validity and reliablity test. Uncovering the results of both previous evaluations, the correlation of the results between each teacher is conducted.

Normal Distribution Test

Using SPSS program, particularly Kolmogorov-Smirnov test, the data is calculated to find out if the distribution is normal. The application of Kolmogorov Smirnov test is that when the significance is below 0.05 means that the tested data is significantly different from standard normal distribution. This also means that the data is not normally distributed. The data is normal when the significance is greater then 0.05. The following is the out put of the data:

NPar Tests

One-Sample Kolmogorov-Smirnov Test

Table 3 - One sample of Kolmogorov-Smirnov Test

		Teacher 1	Teacher 2	Teacher 3
N		21	21	21
Normal Parameters(a,b)	Mean	6.2381	6.4524	6.2381
	Std. Deviation	.56167	.78906	.62488
Most Extreme Differenc	Absolute	.251	.145	.209

es				
	Positive	.178	.145	.172
	Negative	-.251	-.098	-.209
Kolmogorov-Smirnov Z		1.150	.666	.957
Asymp. Sig. (2-tailed)		.142	.767	.320

a Test distribution is Normal.
b Calculated from data.

		Teacher 1	Teacher 2	Teacher 3
Teacher1	Pearson Correlation	1	.619(**)	.721(**)
	Sig. (2-tailed)		.003	.000
	N	21	21	21
Teacher2	Pearson Correlation	.619(**)	1	.886(**)
	Sig. (2-tailed)	.003		.000
	N	21	21	21
Teacher3	Pearson Correlation	.721(**)	.886(**)	1
	Sig. (2-tailed)	.000	.000	
	N	21	21	21

The SPSS out put above can be seen particularly on Asymp.Sig (2-tailed); Teacher 1 is 0.142, Teacher 2 is 0.767 and Teacher 3 is 0.320. It can be interpreted that all the out puts are normal because all the numbers are greater than 0.05. As soon as the standard normal distribution is identified, Validity and relibility of the data can be measured.

Validity and Reliability Test

Validity and reliability test is also done using SPSS. Since the standard normal distribution is confirmed, the test can be done using pearson test. According to commonly known statistic theory, the data is valid when the Cronbach's Alpha shows greater than 0.7. This research focuses one kind of reliability: internal consistency reliability. Internal consistency reliability is used to measure the cosistency of certain rater. The following is the out put of the internal consistency reliability with four speaking assessment criteria, namely fluency and coherence, lexical resources, grammatical range and accuracy, and pronunciation.

Reliability Statistics

Table 4 - Cronbach's Alpha

On the *Reliability Statistics* table out put can be clearly identified that the Cronbach's Alpha is 0.817 and since this one is greater than 0.7, it can be concluded that the data is valid. Also, the score on *Cronbach's Alpha Based on Standardized Items* which shows the reliability of the data as a whole, indicated 0.822, the greater the score the more reliable it is. The Table for N 21 (in table 3.1) can be found that the DF (N-2) for the degree of 0.05 is 0.3687 which can be obviously seen that the data is valid and reliable because 0.817 > 0.7 (valid) and 0.822 > 0.3687 (reliable).

Correlation Test

In this particular paper, finding out correlation in paired items is the final steps. As it is explained in chapter two that standard normal distribution can be tested its correlation using the most common way, Pearson Correlation. It is stated that when the score of significance is positive and below 0.05, the correlation is significant and

linear relationship. It means that as one variable gets larger, the other gets larger in direct proportion.

Correlations

Table 5 - Correlation

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

The table that Shows that "correlation is significant at the 0.01 level (2-tailed), it can be recognized that the pair relation between each teacher's result is significant as it is below 0.05

CONCLUSION

From the very first steps of the process, collecting the data, up to the final steps, identifyng correlation, can be summarized the teachers in Inlingua Internatinal English training produce reliable and valid results of the speaking IELTS subtest, even though the teachers have different working span experience because they use the same rubric for measuring their students' speaking performance. More than that, the correlation between one group on scores to others are significant and linear. This can be reference for the future in the

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.817	.822	4

way Inlingua International maintain their teachers' marking scheme.

REFERENCES

Bussiness Dictionary, Accessed 20 October 2013. http://www.businessdictionary.com/definition/normal-distribution.html

Cambridge, Official Examination papers from University of Cambridge Esol Examination.Cambridge IELTS 8 With Answer

Farlex, Free Dictionary,.http://medical-dictionary.thefreedictionary.com/Correlation+%28in+statistics%29, accessed 20 Oct 2013

IELTS Nine Band Scale. http://www.ieltsessentials.com/results/ielts_9-band_scale.aspx, Accesed on 13 July 2013

Phelan, Collin and Wren, Julie, 2005-2006. Graduate Assistants, UNI Office of Academic Assessment, *Exploraing Reliability in Academic Assessment,* http://www.uni.edu/chfasoa/reliabilityandvalidity.htm. Accesed 20 Oct 2013.

The Survey System.Correlation. http://www.surveysystem.com/correlation.htm, Accessed on 25 August 2013.