International Journal of Innovation Scientific Research and Review
Vol. 03, Issue, 01, pp.688-694, January, 2021
Available online at http://www.journalijisr.com

IJISRR

Research Article

ISSN: 2582-6131

THE IMPORTANT ROLE OF TESTING IN ENGLISH TEACHING AND LEARNING PROCESS IN VIETNAM

*** Vu Viet Phuong MA**

UNETI - Hanoi - Vietnam

Received 25th September 2020; Accepted 15th December 2020; Published online 20th January 2021

ABSTRACT

Testing plays a very important role in the teaching and learning process. It is considered as an integral part of any teaching process. If teaching is a process of helping learners discover "new" ideas and "new" ways of organizing what they learn, testing is an important tool to measure what learners achieve through process of teaching. Therefore, testing makes a remarkable contribution to the success of teaching and learning activities. Furthermore, tests are used not only to evaluate teaching and learning results but also to promote teaching and learning activities in such a way that it helps teacher understand his/her students' ability, interest, attitudes and needs in order to teach and motivate them.

Keywords: testing; teaching and learning process; motivate.

INTRODUCTION

As we know, there are many types of test and each test is used for different purposes. There are four types of test such as proficiency tests, achievement tests, diagnostic tests and placement tests. IELTS means International English Language Testing System and it is a test of English language proficiency. It consists of four language skills - listening, reading, writing and speaking. The listening test which will be evaluated in this assignment is one of the four skills in the Academic module. First, a brief literature review of language testing and evaluation in general and the language knowledge and skills relating to the test to be evaluated will be presented. Next some brief information about the test will be presented regarding the test content, components and structure, timing, and scoring keys. Finally, and most importantly, the validity of the test will be analyzed with regard to test content, test structure, construct validity, and test reliability based on the scoring. The test analysis shows that while more should be put into consideration as to the test content, structure, and scoring keys, the high item discrimination index values suggest high validity and reliability of the test items.

LITERATURE REVIEW

Language testing

Testing is an important part of every teaching and learning experience and becomes one of the main aspects of methodology.
A test, in Caroll's (1968, p.46, cited in Bachman, 1995, p.20) words, is "*a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual.*" Heaton (1990) holds that tests should be considered first as means of assessing the students' performance and then as devices to motivate students. Moss (1998) defines that a test is any procedure for measuring ability, knowledge, or performance. In line with this definition, Brown (2004) also states that a test, in a plain word, is a method of measuring a person's ability or knowledge in a given domain. This proposed definition shows some essential components.

***Corresponding Author: Vu Viet Phuong MA,**
UNETI - Hanoi - Vietnam.

Language testing is one of the forms of testing and it is also one form of measurements. According to McNamara, a language test is "a procedure for gathering evidence of general or specific language abilities from performance on tasks designed to provide a basis for predictions about an individual's use of those abilities in real world contexts" (p. 11). Moreover, Bachman (1990:20) claims that what distinguishes a test from other types of measurement is that it is designed to obtain specific sample of behavior. This distinction is believed to be of great importance because it reflects the primary justification of the use of language tests and has implications for how we design, develop and use them to their best use. Thus, language tests can provide the means for more focus on the specific assure of interest. In sum, a language test is an instrument for assessing test-takers' use of language knowledge and skills for communicative purposes. It can also play the role of a motivating device for students in their learning process and for teachers to adjust their teaching accordingly. In term of purposes, language tests are divided into achievement tests and proficiency tests (Mc Namara, 2000). "Proficiency tests are designed to measure people's ability in a language, regardless of any training they may have had in that language" (Hughes, 2003). They look to the future situation in language use without necessarily any reference to the previous process of teaching (Mc Namara, 2000). "In contrast to proficiency tests, achievement tests are directly related to language courses" (Hughes, 2003); they relate to the past in that they measure what language the students have learned as a result of teaching.

Proficiency test

According to Hughes (1990:9), "Proficiency tests are designed to measure people's ability in a language regardless of any training they may have had in that language." That is to say the content of a proficiency test is not based on the content or objectives of any language course test takers may have followed. It is rather based on a specification of what they have to be able to do in the language to meet the requirement of their future aims. Other test specialists, Harrison (1986) and Henning (1987) share the view that proficiency test helps both teachers and learners know whether the learners can be able to follow a particular course or they have to take some pre - departure training to some other popular tests such as TOEFL, IELTS, which are used to test students' proficiency for their study in

some English speaking countries. In Vietnam proficiency tests are of different levels namely A, B, C for workers, engineers, teachers, architects, etc.

Test evaluation

Test validity

It should be noted that different scholars think of validity in different ways. Heaton (1988: 159) also provides a simple but complete definition of validity as "the validity of a test is the extent to which it measures what it is supposed to measure". Hughes (1989: 22) claimed that "A test is said to be valid if measures accurately what it is intended to measure".

Content validity

Content validity refers to the extent which the test provides both a satisfactory sample of the syllabus and information about the students' ability in the aspects we are interested in. Content validity is very important in evaluating the validity of the test in terms of that "the greater a test's content validity, the more likely it is to be an accurate measure of what is supposed to measure" (Hughes, 1989: 22). Content of a test is determined by what is easy to test rather than what is important to test.

Construct validity

Construct validity is viewed from a purely statistical perspective in much of the recent American literature Bachman and Palmer (1981a). To understand whether a piece of research has construct validity, three steps should be followed. First, the theoretical relationships must be specified. Second, the empirical relationships between the measures of the concepts must be examined. Third, the empirical evidence must be interpreted in terms of how it clarifies the construct validity of the particular measure being tested (Carmines & Zeller, 1991: 23).

Face validity

A test is said to have face validity if it looks as if it measures what it is supposed to measure. Face validity is very closely related to content validity. While content validity depends on a theoretical basis for assuming if a test is assessing all domains of a certain criterion, face validity relates to whether a test appears to be good measure or not.

Criterion-related validity

Criterion-related validity is used to demonstrate the accuracy of a measure or procedure by comparing it with another measure or procedure which has been demonstrated to be valid. Criterion-related validity consists of two types (Davies, 1977): concurrent validity, where the test scores are correlated with another measure of performance, usually an older established test, taken at the same time (Kelly, 1978; Davies, 1983) and predicative validity, where test scores are correlated with some future criterion of performance (Bachman and Palmer, 1981).

Test reliability

According to McNamara, reliability is consistency of measurement of individuals by a test, usually expressed the extent to which individuals have been measured consistently by a test. Three aspects of reliability are usually taken into account. The first concerns the consistency of scoring among different markers. The second is the concern of the tester how to enhance the agreement between markers by establishing, and maintaining adherence to, explicit guidelines for the conduct of this marking. The third aspect of reliability is that of parallel - forms of a test to be devised. Rayan (2002) defines reliability as "the extent to which results are consistent over time and an accurate representation of the total population under study is referred to as reliability and if the results of a study can be reproduced under similar methodology, then the research instrument is considered to be reliable." Kirk and Miller (1986) identify three types of reliability: (1) the degree to which a measurement, given repeatedly, remains the same; (2) the stability of a measurement over time; and (3) the similarity of a measurement within a given time period. The present assignment will look at the valid and reliable scoring of an IELTS test regarding the language skills of listening. For Kitao and Kitao (1996b, p.1), testing listening is challenging as it is difficult to design tests that reflect real-world listening tasks. In order to test listeners' understanding of meaning, Kitao and Kitao (1996b, pp. 3-6) suggests the following test tasks:

Understanding sentences and dialogues:

- Interpreting meaning: Listeners interpret the meaning of certain heard utterances or strings of utterances.
- Responding to utterances: Listeners choose the correct responses to the heard utterances. This is considered a more communicative type of task than many other listening tasks.

Task using visual materials

- Matching and True/False tasks: Tastes look at visual materials (i.e. pictures, charts, graphs, etc.) and match them with spoken statements, dialogues or descriptions. An alternative is to look at the visual material and decide whether the spoken statement or description is true or false.
- Mapping tasks: Tastes listen for the directions to somewhere and follow the map.
- Drawing tasks: Tastes listen to instructions and do drawing tasks.

Tasks involving talks and lectures

- Summary - filling/ Table completion
- Short answers/ Sentence completion/ Multiple choice/ True-False-Not Given
- Taking notes and using notes to answer questions

TEST INFORMATION (SPECIFICATIONS)

Appendix 1 appearing at the end of this assignment is a copy of the test being evaluated, which should provide sufficient information about the test content, test operations – skills and language elements to be tested, and test structure. More details and explanations about the test are presented below.

Kind of test

IELTS means International English Language Testing System and it is a test of English language proficiency. It covers all four language skills - listening, reading, writing and speaking. There are two IELTS modules including Academic and General Training. The listening test which will be evaluated in this assignment is one of the four skills in the Academic module.

Test purposes

The purpose of the IELTS Listening Module is to:

- Assess the language ability of candidates who want to study or work where English is the language of communication
- Establish your ability to function on a daily basis in a country where English is spoken as a first language;
- Establish your ability to function in an academic environment where English is used as a tuition medium.

Test takers

Many English language learners need IELTS in order to pursue academic or non-academic training. IELTS is used by educational institutions, governments, professional bodies and commercial organizations. The General Training module is for candidate wishing to migrate to an English - speaking country and for those wishing to train or study at below degree level. The Academic module evaluated below is for candidates wishing to study at undergraduate or post graduate levels, and for those seeking professional registration.

Abilities to be tested

There are 40 questions. A variety of question types is used concluding multiple choice, matching, plan/map/diagram labeling, form completion, note completion, table completion, short- answer questions, and summary completion.

Question type	Skills
Summary completion	Predicting
Sentence completion	Listening for gist
Gap-fill questions	Listening for specific information
	Listening for main ideas
	Listening for and recognizing signpost words
	Understanding a speaker's attitude/opinion
	Recognizing the speaker's role
	Recognizing speakers' pronunciation
	Focusing on more than one question at a time
Completing notes	Predicting
Short answers	Listening for specific information
	Listening for gist
	Listening for and recognizing signpost words
	Listening for main ideas
	Understanding a speaker's attitude/opinion
	Recognizing the speaker's role
	Focusing on more than one question at a time
	Guessing the meaning of unfamiliar words
Diagram Labeling, namely:	Predicting
flow chart	Listening for gist
map	Listening for specific information
process	Listening for and recognizing signpost words
picture of an object	Listening for main ideas
	Understanding a speaker's attitude/opinion
	Recognizing the speaker's role
	Recognizing speakers' pronunciation
	Focusing on more than one question at a time
Table completion	Predicting
Chart completion	Listening for gist
Grid completion	Listening for specific information
	Listening for and recognizing signpost words
	Listening for main ideas
	Understanding a speaker's attitude/opinion
	Recognizing speakers' pronunciation
	Recognizing the speaker's role
	Focusing on more than one question at a time
	Listening for comparisons
Classification	Predicting
Matching	Listening for specific information
	Listening for and recognizing signpost words
	Listening for main ideas
	Understanding a speaker's attitude/opinion
	Recognizing the speaker's role
	Focusing on more than one question at a time

Use of the test result

The result of the test is used to:

- Relieve the institution of all the administrations and cost involved in English language testing
- Select candidates who already meet the English language requirements
- Gain access to ongoing support from some of the world's leading language assessment experts.
- Provide the applicants with a clearer understanding of the level of English they need.

TEST VALIDITY ANALYSIS

Content validity

Operation Ability to catch key words and obtain the gist
Type of text: Dialogue (Section 1, 3), Monologue (Section 2, 4)
Section 1: Dialogue: conversation
Section 2: Monologue: message giving information
Section 3: Dialogue: conversation
Section 4: Monologue: Lecture
Length: ***Section 1***: 7, 10 minutes
Section 2: 6, 39 minutes
Section 3: 7, 12 minutes
Section 4: 6, 31 minutes
Topics:

Section 1: a conversation between a clerk at the enquiries desk of a transport company and a man who is asking for travel information.
Section 2: a guidance counselor talking to a group of students
Section 3: a conversation between a tutor and two students who are preparing for an English literature test.
Section 4: a talk on the topic of time perspectives

Structure, timing, medium and techniques

Test structure
There are 4 separate sections in Listening test. In each section, there are 10 questions.

Number of items
In total, there are 40 items

Timing
40 minutes in total. Approximately 30 minutes of listening and 10 minutes of transferring answer onto the answer sheet.

Medium
Pencil, paper, CD player and headphones

Criteria levels of performance
Scores 0-14
Unlikely to get an acceptable score on the IELTS listening test under examination conditions and need to spend a lot of time improving your English before you take IELTS.
Scores 15-25
May get an acceptable score on the IELTS
Scores 26 and above
Likely to get an acceptable score on the IELTS listening test under examination conditions

Scores 26- 27: 6.0 IELTS Band Score, Competent User

Scoring Procedure

Answer will be written on a separate sheet, key is given so clerical staff can mark the test easily.

Score	IELTS Band Score	Description	Detailed Description
26-27	6.0	Competent User	Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use and understand fairly complex language, particularly in familiar situations.
30-32	7.0	Good User	Has operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning.
35-36	8.0	Very Good User	Has fully operational command of the language with only occasional unsystematically in accuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.
39-40	9.0	Expert User	Has fully operational command of the language: appropriate, accurate and fluent with complete understanding.

* Source: http://www.ielts.org

Score distribution

Score	Proportion of students
33-40	0%
30-32	0%
28-29	0%
26-27	22%
24-25	11%
22-23	11%
19-21	22%
17-18	11%
15-16	0%

It is obvious from the table that most of the students tested get from 26 to 27 (equivalent to 6.0) and from 19 to 21 (5.5) correct answers. No students get more than 32 as well as get under 16 correct answers.

Difficulty level and discrimination level

Item	Right	Wrong	%Right	%Wrong
Section 1				
1	7	2	78%	22%
2	6	3	67%	33%
3	4	5	44%	56%
4	1	8	11%	89%
5	5	4	56%	44%
6	5	4	56%	44%
7	5	4	56%	44%
8	7	2	78%	22%
9	1	8	11%	89%
10	2	7	22%	78%
Section 2				
11	4	5	44%	56%
12	7	2	78%	22%
13	8	1	89%	11%
14	2	7	22%	78%
15	5	4	56%	44%
16	5	4	56%	44%
17	8	1	89%	11%
18	2	7	22%	78%
19	3	6	33%	67%
20	1	8	11%	89%
Section 3				
21	5	4	56%	44%
22	6	3	67%	33%
23	2	7	22%	78%
24	1	8	11%	89%
25	7	2	78%	22%
26	0	9	0%	100%
27	2	7	22%	78%
28	8	1	89%	11%
29	4	5	44%	56%
30	4	5	44%	56%
Section 4				
31	7	2	78%	22%
32	3	6	33%	67%
33	9	0	100%	0%
34	9	0	100%	0%
35	2	7	22%	78%
36	3	6	33%	67%
37	2	7	22%	78%
38	6	3	67%	33%
39	1	8	11%	89%
40	3	6	33%	67%

This is a typical IELTS listening test used in an international exam with 40 items distributed equally in 4 separate sections. Multiple-choice items account for 16 of 40 ones. The rest are gap-filling items, which requires test-takers listening tape scripts and writing answers for gaps. So, key words must be present in right places with appropriate space and must not lie too closely with each other. There is also one only marking skill for gap-filling answers: only completely correct answers could be accepted as scoring. Partly or half correct answers will be considered the same as wrong ones. The use of the test result is used to judge whether or not a test taker is eligible to be admitted to a university in English-speaking countries. In another word, the interpretation of test results by receiving institutions involves relating course requirements to the candidate's proficiency in English as indicated by the Overall Band Score and by individual sub-test scores. The appropriate level required for a given course is ultimately something which institutions/faculties/departments/course tutors must determine in the light of knowledge of their own courses and their experience of overseas students taking them. In some departments, the nature of the work makes fewer demands on language competence than in others and students can cope with a relatively low level of English. On the other hand, some fields require particularly sophisticated language skills: some make higher demands on students' reading and writing skills and others require advanced oral presentation skills. Such factors naturally have to be taken into account when deciding whether or not further language instruction is required.

Difficulty level:

According to McNamara (2000, p.60), item facility or item difficulty expresses the proportion of the people taking the test who got the item right. The difficulty of an item is understood as the proportion of the persons who answer a test item correctly. The higher this proportion is, the lower the difficulty. What this means is that it has to do with an inverse relationship: the greater the difficulty of an item, the lower its index (Wood, 1960). To calculate the difficulty of an item, the number of persons who answered it correctly is divided by the total number of the persons who answered it. As we can see from the chart above, the percentage of the test takers who got the right answer can be translated into the proportion. Usually, this proportion is indicated by the letter *p,* which implies the difficulty of the item (Crocker and Algina, 1986). It is calculated by the following formula:

$$P(i) = A(i) / N(i)$$

In which:

P(i) = Difficulty index of item *i*
A(*i*) = Number of correct answers to item *i*
N(*i*) = Number of correct answers plus number of incorrect answers to item *i*
For instance, with item no.1, we have 7 over 9 got the right answer = 78% => the proportion is 0.78 (7 divided by 9). This test is not only a proficiency test but also a criterion-referenced test. The ideal item difficulty is 0.5 but of course, it is hard to hit this target exactly, and a range of item facilities from 0.33 to 0.67 is usually accepted (McNamara, 2000, p.61).
*Note:
High item difficulty = H
Average item difficulty = A
Low item difficulty = L

Section 1

Item no.01: P = 0.78: H

Item no.02: P = 0.67: A
Item no.03: P = 0.44: A
Item no.04: P = 0.11: L
Item no.05: P = 0.56: A
Item no.06: P = 0.56: A
Item no.07: P = 0.56: A
Item no.08: P = 0.78: H
Item no.09: P = 0.11: L
Item no.10: P = 0.22: L

Section 2

Item no.11: P = 0.44: A
Item no.12: P = 0.78: H
Item no.13: P = 0.89: H
Item no.14: P = 0.22: L
Item no.15: P = 0.56: A
Item no.16: P = 0.56: A
Item no.17: P = 0.89: H
Item no.18: P = 0.22: L
Item no.19: P = 0.33: A
Item no.20: P = 0.11: L

Section 3

Item no.21: P = 0.56: A
Item no.22: P = 0.67: A
Item no.23: P = 0.22: L
Item no.24: P = 0.11: L
Item no.25: P = 0.78: H
Item no.26: P = 0: no-one got the item right (very difficult)
Item no.27: P = 0.22: L
Item no.28: P = 0.89: H
Item no.29: P = 0.44: A
Item no.30: P = 0.44: A

Section 4

Item no.31: P = 0.78: H
Item no.32: P = 0.33: A
Item no.33: P = 1: everybody got the item right (very easy)
Item no.34: P = 1: everybody got the item right (very easy)
Item no.35: P = 0.22: L
Item no.36: P = 0.33: A
Item no.37: P = 0.22: L
Item no.38: P = 0.67: A
Item no.39: P = 0.11: L
Item no.40: P = 0.33: A

After calculating, we have this result:
28% of all items got High Item Difficulty
42.5% of all items got Average Item Difficulty
32.5% of all items got Low Item Difficulty

As we can see from the result above, the majority of all items are at Average difficulty level; so the test is quite appropriate and relevant. According to McNamara (2000, p.61), items with high item difficulty may be useful to include some at the beginning of a test in order to ease candidates into the test and to allow them a chance to get over their nerves. Most items of Section 1 (supposed to be the easiest section out of 4) satisfy this point. Moreover, to see it clearer, we can analyze some specific cases:

- Item no.04: P = 0.11.This is very low item difficulty. Although the question is not too hard, most test takers failed to have correct answer (Ans: A21 to (A) 24). Some of them write the letter H instead of A (quite similar pronunciation) and some of them just wrote the numbers. Both cases led to their missing points. The low item difficulty in this item is mainly because the test takers cannot hear clearly.
- Item no.09: P = 0.11. This is very low item difficulty. The question requested an answer of a series of words and numbers, and again, most test takers failed to write them correctly. The common mistake among them is the mis-distinguishing of some words or numbers which have quite similar pronunciation (A-H, G-J, H-8…). The low item difficulty in this item is mainly because the test takers cannot hear clearly.
- Item no.13: P = 0.89. This is very high item difficulty. This is easy to understand because the information needed to complete the question is easy to hear and locate.
- Item no.20: P = 0.11. This is very low item difficulty. A majority of test takers failed to answer the question because of the different expressions between the listening script and the clues given on the testing paper, so, test takers couldn't locate the information needed.
- Item no.24: P = 0.11. This is very low item difficulty. The problems here maybe because the expression used in the listening typescript is too difficult compared to one used in the testing paper; so, test takers were confused and couldn't get the right answer.
- Item no.26: P = 0. This means that the item is too hard (no-one got the item right) and it doesn't help us to distinguish between the test takers. This result is because of the too difficult vocabulary used in the listening script.
- Item no.33: P = 1. This means that the item is quite easy (everybody got the item right). This is quite easy to understand because test takers got a clear clue on the testing paper, so it was easy for them to locate the information needed.
- Item no.34: P = 1. This means that the item is too easy (everybody got the item right). This is quite easy to understand because test takers got a clear clue on the testing paper), so it was easy for them to locate the information needed.
- Item no.39: P = 0.11. This very low item difficulty. This is mainly because there is too much distracting information in the listening script that makes it difficult for test takers to have right answer.

According to the detailed analysis on some special cases above, we can see that, in general, the test was designed with the quite appropriate difficulty level. However, for items like no.33 and no.34, the result was not quite persuasive because it is not supposed to have questions that all test takers can do right at the end of the test – where we expect to have more distinguishing questions.

Discrimination level

The index of discrimination is a numerical indicator of how the poorer students answered the item as compared to how the better students answered the item. The scores are divided into three groups with the top 1/3 of the scores in the upper group and the bottom 1/3 in the lower group. The number of correct responses for an item by the lower group is subtracted from the number of correct responses for the item in the upper group. The difference in the number correct is divided by the number of students in either group. The process is repeated for each item.

Example: 9 students take a test. The top 3 scores and the bottom 3 scores are the upper and lower groups. For item no. 1, all 3 students

in the upper group answered the item correctly while 1 student in the lower group answered correctly. The index of discrimination (D) for item no. 1 would be calculated as follows:

$$D = \frac{3-1}{3} = 0.67$$

We have the following discrimination indices:

Item number	Discrimination index (D)
1	0.667
2	0
3	1
4	0.333
5	0
6	0.333
7	0
8	-0.333
9	0
10	0.667
11	0.667
12	0.667
13	0
14	0.333
15	0
16	0.667
17	0.667
18	0.667
19	0.333
20	0.333
21	0
22	0
23	0.333
24	-0.333
25	0
26	1
27	0.333
28	0.333
29	1
30	0.667
31	0.667
32	0
33	0.333
34	0
35	0.667
36	-0.333
37	0.333
38	0.667
39	0.667
40	0.333

Ebel and Frisbie (1986) give us the following rule of thumb for determining the quality of the items, in terms of the discrimination index. The table below shows the values D and their corresponding interpretation. The recommendations for each of these values are shown in the table as well.

Discrimination power of the answers according to their D value

D =	Quality	Recommendations
> 0.39	Excellent	Retain
0.30 - 0.39	Good	Possibilities for improvement
0.20 - 0.29	Mediocre	Need to check/review
0.00 - 0.20	Poor	Discard or review in depth
< -0.01	Worst	Definitely discard

For a small group of students like in this test, an item discrimination index that exceeds 0.20 is considered satisfactory. For larger groups, the index should be higher because more difference between groups would be expected.

As for this point, we have this result:

- 3 items (7.5%) discriminate negatively (<0)
- 11 items (27.5%) discriminate poorly (0.00 – 0.20)
- 11 items (27.5%) discriminate well (0.30 – 0.39)
- 12 items (30%) discriminate excellently (>0.39)
- 3 items (7.5%) got ideal discrimination index (1)

According to the table above, there are a lot of items with problems of discrimination (35%); and theoretically it means that the coming out of the test is very confusing and not reliable. But this test included a very small number of candidates (say 9), so formal discrimination indices calculated as above are not very meaningful. It is just worthwhile dividing the students into two groups – top half and bottom half – and then comparing their performance on each item (Hughes, 2003, p.228). Then the guidelines for an acceptable level of discrimination depend upon item difficulty. For very easy or very difficult items, low discrimination levels would be expected; most students, regardless of ability, would get the item correct or incorrect as the case may be. For items with a difficulty level of about 70 percent (0.70), the discrimination should be at least 0.30. Accordingly, we have the analysis of some special cases as mentioned in Part 1 (difficulty level):

- Item no.04: P = 0.11 (very difficult), D = 0.667 (high)
- Item no.09: P = 0.11 (very difficult), D = 0 (very low)
- Item no.13: P = 0.89 (very easy), D = 0 (very low)
- Item no.20: P = 0.11 (very difficult), D = 0.33 (average)
- Item no.24: P = 0.11 (very difficult), D = -0.33 (very low)
- Item no.26: P = 0 (very difficult), D = 1 (very high)
- Item no.33: P = 1 (very easy), D = 0.33 (average)
- Item no.34: P = 1 (very easy). D = 0 (very low)
- Item no.39: P = 0.11 (very difficult), D = 0.67 (high)

Effective items should discriminate well between high and low scoring candidates. In the cases analyzed above, the items no.13 and no.34 are too easy to discriminate. Items no.09 and no.24 is too difficult to discriminate. Items no.04, no.26 and no.39 are the items with high difficulty level, but their high discrimination level in such cases is reasonable because items no.26 and no.39 are used to discriminate between the strongest test takers; while item no.04 requires very careful listening to get right answer. With items no.20 and no.33, the discrimination level is average although the items are either very difficult or very easy. The data analysis above can help us to conclude to some extent about the reliability of this test. In general, the difficulty level and discrimination level reflect quite appropriate and relevant test items. The scores of the test provide quite reliable information on test takers' abilities. However, the test was conducted with a small number of candidates and not too many test items, so the reliability based on discrimination index is not very meaningful. Due to the lack of condition, we cannot have exact reliability coefficient. Based on the item difficulty and discrimination indices, a majority of items satisfy; however, there are still some items where there is no difference between the groups or reflect wrongly the test takers' ability overall, then these items need scrutinizing.

Comments on test validity

This is a proficiency test "designed to measure people's ability in a language regardless of any training they may have had in that language" (Hughes, p. 9). In IELTS examinations:

- The Academic Version intended for those who want to enrol in universities and other institutions of higher education and for professionals such as medical doctors and nurses who want to study or practice in an English-speaking country.

- The General Training Version is intended for those planning to undertake non-academic training or to gain work experience, or for immigration purposes.

That's why this kind of test is very different for others that "are associated with the process of instruction" - achievement tests (Hughes), "are designed to identify students' strengths and weaknesses to ascertain what further teaching is necessary" - diagnostic tests (Hughes), or placement tests providing "information which helps to place students at the stage or the teaching program most appropriate to their abilities" (Hughes). In the IELTS listening above, the gap-filling form questions account for 70% (equivalent to 24 of 40 items) and the rest of 30% is multiple choice questions. However, the distribution among them is not equal for each section (there are four sections and each includes 10 items). In Section 1, there are no multiple choice questions. In the next section, multiple choice questions also witness equal in number and then the kind of question reduce sharply when there are only two in the final section. This is also the most challenging for all candidates when they have to perform a very good ability of language to successfully do the task.

CONCLUSION

The IELTS test above has shown it's an internationally standard test with many good points. First, it directly asks what is intended to be asked. All keys are very clear, and they all contain important information. Second, all topics selected in this test are very familiar. Words and structure ranges are of daily matters, and lecturing, easy to understand and contain no especially cultural phrases. International test takers can understand the contents and meet little difficulty in perceiving the meanings of words and sentences. Furthermore, all instructions are explicit and the form of the test is also very common and popular with every test taker.

However, it still generates some weak points such as the contribution of type of questions in each section; some items have very poor discrimination level. In short, this is a test with high reliability and can satisfy international requirements in testing contestants' ability in English.

REFERENCES

1. Alderson, J.C., Claphan, C., Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: Cambridge University Press.
2. Bachman, L.F. (1990). Fundermental Considerations in Language Testing. Oxford: Oxford University Press.
3. Bachman, L.F., Palmer, A.S. (1996). Language Testing in Practice. London: Oxford
4. Davies, A. (1996). Language testing. Cambridge: Cambridge University Press..
5. Heaton, J.B.(1998). Writing English Language Tests. London: Longman Group UK, Ltd.
6. Henning, G. (1989). A Guide to Language Testing. Cambridge: Newbury House Publishers.
7. Hughes, A. (1989). Testing for Language Teachers. Cambridge: Cambridge University Press.
8. Lado,R. (1961). Language Testing. London: Longman
9. McNamara, T. (2000). Language Testing. Oxford: Oxford University Press.
10. Palmer, A.S., and L.F. Bachman, (1981). Basic concerns in test validation. (in) Alderson, J.C. and A. Hughes (eds.), (1981)
11. Official IELTS Practice Materials (2010). Retrieved 5th April 20112 from http://www.ielts.org, http://takeielts.britishcouncil.org/sites/default/files/Listening_practice_questions_121012.pdf

****