

## Assessing the Difficulty Levels of Different Test Types in Listening Tests for Second Year English Specialization Students at Dagon University

Tin Win Yee\*

### Abstract

This paper focuses on a study which investigated difficulty levels in listening tests and student's performance in those tests. To this end, a comparison was made between the results of six listening tests. Different linguists formulated different models and methods in assessing the various tests. A statistical model used in this research is modified, based on J.B. Heatson (1988) and Alan Graham (1999). Through the comparative analyses, student's perceived task and test difficulties in the listening tests were found to correlate, to a great extent, with the results of question analysis. Implications are drawn for test administrators of listening tests and language teachers.

### Introduction

Examination or testing is an integral part of a teaching-learning process which allows teachers to evaluate their students during and at the end of an educational course. Many teachers dislike preparing and assessing scores and most students dread doing exams. Tests are powerful educational tools which measure students' ability in their respective course. When we learn a language, there are four skills that we need for complete communication. Speaking and writing are regarded as productive skills and listening and reading are regarded as receptive skills. Receptive skills are the ways in which people extract meaning from the discourse they see or hear. There are generalities about this kind of processing which apply to both reading and listening but there are also significant between reading and listening process. Most of the students in Myanmar said that listening test is the most difficult in second language learning. Listening involves real-time processing, generally without the option of going back to earlier sections of the text the students may have missed (Buck, 2001; Flowerdew, 1994). The students cannot get a second chance to catch the information. They typically encounter disfluencies, false start and pronunciation.

This paper aims to investigate difficulty levels of different listening test types regarding scores of second year English specialization students at Dagon University.

The objectives of this paper are:

1. to identify difficulty level of each listening test
2. to find out discrimination index of each item in six listening tests
3. to investigate correlation between difficulty levels and discrimination indexes related to students' score in listening comprehension tests and
4. to suggest teachers how to develop good tests.

Testing is a process which involves the adoption, development and adaption of instruments of educational measurement. The principle function of an instrument of educational measurement, when the instrument is created as a means of inferring students' abilities, is to offer information on how to make correct decisions on formulating test items. In this paper, correlation between difficulty levels and discrimination indexes of six tests is calculated in order to give suggestions to teachers.

Based on the aim of the paper, the following research questions are adopted:

1. According to Harold S. Madsen's model, is difficulty level of each test too easy, easy, difficult or too difficult?

---

\*Tutor, Department of English, Dagon University

2. What index is represented for discrimination of each test?
3. What kinds of correlation can be found out between difficulty levels and discrimination indexes?

### **Literature Review**

#### **A Short History of Language Testing**

The development of modern linguistic theory has helped to make language teachers and testers aware of the importance of assessing language tests. Spolsky (1978) and Hinofotis (1981) cited by James Dean Brown, both have pointed out that language testing can be broken into periods, or trends, of development. Hinofotis labeled them the prescientific movements, the psychometric/structuralist movement, the integrative/sociolinguistic movement and communicative movement.

The prescientific movement (before 1939) is characterized by translation and free composition tests developed by the classroom teachers when it comes to developing tests. These types of tests are difficult to score objectively. The psychometric-structuralist movement (1950s-1960s) arise the concepts objectivity, reliability and validity. So, language tests became increasingly scientific, reliable and precise. The tests, usually in multiple-choice format, are easy to administer and score. This movement is important because language test development follows scientific principles as the first step. The integrative-sociolinguistic movement (after 1960s) has its roots in the argument that language is creative. The development of communicative competence depended on more than simple grammatical control of the language. Communicative competence involves knowledge of the language appropriate for different situations. Tests typical of this movement were the cloze test and dictation, both of which assess the students' ability to manipulate language written a context of extended text. This usher students to developing communicative skills related to critical thinking and problem-solving, initiative, effective oral and written communication, synthesizing and analyzing information, and curiosity and imagination skills (Dr. Wagner).

#### **Two Skills-Based Issues**

Language tests are often separated into skill areas: reading, writing, listening and speaking. The two skills-based issues involve in these areas. Language teachers and testers usually use the terms "the channel issue" and "the mode issue". The channel issue is divided into two categories: written channel and oral channel. Reading and writing tests can be referred as written channel subtests because they both involve language written on paper. Listening and speaking tests can be labeled oral channel subtests because they involve the use of sound to communicate. For the mode issue, the production mode and the receptive mode include. The productive mode consists of the skills which send and give information to others – that is writing and speaking. The receptive mode consists of the skills which receive and understand the message from others – that is reading and listening. In language testing, teachers must realize that certain types of test question are more closely related to testing the receptive skills and others are more closely associated with productive skills. Every test necessarily applies at least one channel and one mode at any given time. For example, reading comprehension tests are typically viewed as receptive mode tests of written channel. Composition tests also involve the written channel, but they are in the productive mode.

*Channels*

Written	Oral
Reading Comprehension	Listening Comprehension
Composition	Speaking

**Listening Test Types**

Selecting effective tests to match the purposes of a particular language skill is very important. For testing listening, Authur Hughes (2003: Page.165) describes some useful types of test such as multiple choice, short answer, gap-filling, note-taking and dictation. Some of these tests are use in this paper. Another linguist, Rebecca M. Valette (1977: Page. 76) also mentions different types of listening tests, for example, discrimination of sounds, stress and accent tests, conversation-type tests and listening comprehension, etc. And J.B. Heaton (1988: Page.65) presents some kinds of listening tests; phoneme discrimination tests, tests of stress and intonation, statements and dialogues and passage comprehension. Almost all tests are nearly the same in developing listening test for students.

Most of the language teachers choose carefully the tests for their students in order to get accurate results. Using tests with the wrong types of students can result in mismatches between the tests and the abilities of the students. Sometimes teachers need to modify or edit their adopted tests. If the teachers develop the language tests themselves, the tests should be fit to the goals of the program and to the ability levels and needs of the students. One point to note is that test development is hard work and it can be time-consuming. Even if the tests are complete, it takes time to adapt with the needs of the students.

In this paper, there are six types of test which deals with three levels of listening test. The first one is word level, the second is phrase level and the last is sentence level. Each level uses two different test types.

Multiple choice and gap-filling test types are used for word level tests. In this level, the students need to try to catch up only the words. So this kind of test is the easiest one for the students and most of them can do very well in their tests. Multiple choice items are used to test a wide range of skills. They may require the students to have a detailed understanding of specific points or an overall understanding of the main points of the listening text. There is a question or a sentence beginning followed by three or four possible answers or sentence endings. The students have to choose the one correct answer A, B or C / A, B, C or D.

Gap-filling type of test is a kind of interesting and easy to use to check the students listening skill. Moreover, it focuses on the specific points or information of the listening text. The students need to pay concentration and have to follow the listening text to find the information they need. Gap-filling test is similar to cloze technique. The students need to fill in the gaps and also need lexical knowledge to fit it in the dialogue.

Table completion and gap-filling test types are used for phrase level tests. In this level, the students follow to catch up the phrases or more than one word. So, they may find some difficulties. If the listener doesn't pay attention adequately, the information won't be processed.

Table completion will test students' ability to recognize relationships and connection between facts in the listening text and is most often used with texts dealing with factual information. The students have to complete the list of data or information from the listening text to the table. This test is also measuring the ability to understand spoken English.

Gap-filling test of word level is the same as phrase level. But in phrase level, the students need to be more careful and they should apply their linguistic knowledge and non-linguistic knowledge effectively. Gap-filling can also be used to test a variety of area such vocabulary, grammar and etc. This kind of test type is very useful at testing listening for specific words and phrases.

Answering the questions and dictation test types are used for sentence level tests. This level is the most difficult one for the test-takers and most of the students cannot do very well in this level. They find a lot of difficulties and they may get depressed.

This type of test focuses on a wide variety of skills: understanding overall meaning of the text and making inferences. Answering the questions test takes more time than the others. And the students may find difficulty in inferring the messages. Moreover, they require the ability to construct their responses relatively, quickly and effectively. If the students have enough exposure, they will be able to do well in their listening tests. Another important point is that the questions should be given before they listen, so they have a clear listening purpose.

Dictation is one of the difficult tasks for the students and it assesses performance at all stages of the speech perception process. The basic idea of dictation is simple; test-takers listen to an audio text and write down exactly what they have heard. In this kind of test, the students have to increase the amount of attention and they try to catch all the words spoken by the native speaker. It is true that highly concentrated students do better ones without any concentration. As usual, the students listen to the transcription exercises for dictation. But if the test is leveled up, the material such as the story, the broadcasting news can also be used.

Good evaluation of tests can help the teachers measure student skills more accurately. All listening tests can be improved by taking time to evaluate individual. J.B. Heaton proposed his statistical approach "Writing English Language Tests" in 1988. He shows the ways of checking an individual item which is called item analysis. Item is the smallest unit of every test and item analysis is the process of collecting, summarizing and using information from students' responses to assess the quality of test items. It can result the level of difficulty for each item. Basically, item difficulty and discrimination tells the teachers that how difficult each test is and the difference between high and low students. Then, Pearson's coefficient presented by Alan Graham (1999) is used to show a correlation between the degree of difficulty and the discrimination index.

Teaching listening skill is one of the most difficult tasks for any English language teacher because successful listening skills are acquired over time and with lots of practice (River, 1992). Learning listening skills is frustrating for students because there are no rules as in grammar teaching. Now, assessing becomes essential in language learning. Not only how to assess the tests but also how to design listening assessment procedures, how to validate and evaluate listening tests are involved in assessing listening. It is used in order to help the students to improve listening skills and the analysis of learners' assessment can offer some useful tips in teaching listening skills.

## Research Methodology

### Materials

The materials used in this paper were taken from the course book of New Interchange 3, the story book and some IELTS tests. The different types of listening tests used in this paper were multiple choice, gap-filling, table completion, answering the questions and dictation. Mostly, the topics chosen for the tests were about telling story, radio/TV program, news, advertisements, announcements, interviews and everyday conversations.

### Participants

In this paper, forty students from second year, English specialization classes in 2012-2013 academic year were selected as participants to get scores for item analysis.

### Procedures

The steps in the listening test were as follows:

1. According to the test level, the materials and suitable test types were selected.
2. Then, the teacher introduced the students the topic and elicits some questions to apply their prior knowledge.
3. The teacher gave more detail instructions about the test and sets time allocation.
4. The teacher asked students to review test questions before listening.
5. The teacher played CDs and students listened to it.
6. The listening test was done.
7. After the test, the teacher checked the answers.
8. Finally, the teacher pointed out what types of items students make mistakes and explained how to overcome these difficulties, giving feedback.
9. All six tests were given according to these procedures.
10. The results for different test types were compared for assessment.

### Model Used in the Study

Selection of appropriate question is not enough by itself to ensure a good test. Each question needs to function properly; otherwise, it can affect the tests. So the teachers evaluate the difficulty levels of each test by using different models and methods. In this paper, a simple model presented by J.B.Heaton (1988) is used. It can be called item analysis. And it also helps to find the level of difficulty for every test. According to Harold S.Madsen, a test question is generally considered too easy if more than 90 percent get it right and it is considered too difficult if fewer than 30 percent get it right. It is recorded as easy level if the result is between 90 percent and 60 percent. And it is recorded as difficult level if the percentage is between 60 and 30.

The formula to find difficulty level is as follow:

$$FV = \frac{\text{Correct U} + \text{Correct L}}{2n}$$

FV = facility value or index of difficulty

Correct U = the numbers of student with the correct answer from upper half

Correct L = the numbers of student with the correct answer from lower half

n = the numbers of student in one group

The formula to find discrimination index is as follow:

$$D = \frac{\text{Correct U} - \text{Correct L}}{n}$$

D = Discrimination index

Correct U = the numbers of student with the correct answer from upper half

Correct L = the numbers of student with the correct answer from lower half

n = the numbers of student in one group

Finally, the model formulated by Alan Graham (1999) is used to indicate a definite correlation between the degree of difficulty and the discrimination index. It is known as Pearson's coefficient or Product-moment correlation coefficient.

The formula to find Pearson's coefficient is as follow:

$$r = \frac{S_{XY}}{S_X S_Y}$$

where X is the variable on the horizontal axis and Y is the variable on the vertical axis

$$S_{XY} \text{ is given by the formula } = \frac{\sum XY}{n} - XY$$

$$S_X \text{ is the standard deviation of } X = \sqrt{\frac{\sum X^2}{n} - X^2}$$

$$Y \text{ is the standard deviation of } Y = \sqrt{\frac{\sum Y^2}{n} - Y^2}$$

### Data Analysis

It can be difficult for the students to overcome difficulties in learning a foreign language due to lack of exposure. Although there are numerous factors influencing more or less related to language proficiency, the literature on second language learning proficiency focuses on three types of knowledge – vocabulary size, phonological and grammatical information and background knowledge.

#### Vocabulary size

An obvious factor that can influence comprehension of a spoken text is the overlap between the student's vocabulary knowledge and the vocabulary of the text. Bloomfield (2010) said that students must have an adequate vocabulary to understand the text. Adequate vocabulary might be estimated by the number of words a student needs to know to understand a representative sample of texts. The 5,000 most frequent words give a coverage of 90 to 95 percent of the word in an average passage in many languages. Furthermore, Hirsh and Nation (1992), cited by Bloomfield, have argued that in order to understand all the main points in a text, readers need to be familiar with 95 percent of the words therein. Vocabulary size is an indirect measure of other variables also known to influence listening comprehension ability, including world knowledge. In this paper, about 4,500 frequent words are used for all listening tests.

#### Phonological and grammatical information

Phonological and grammatical information in spoken language comprehension has focused on whether high-ability and low-ability students use top-down and bottom-up processes differently. In specific contexts, both first language and second language students will use top-down processing strategies such as inferencing and elaboration. In fact, the students who relied most heavily on bottom-up information made the most errors (Voss, 1984).

Less experienced second language learning students rely even more heavily on top-down information because they are less familiar with the non-native phonology, vocabulary and syntax. Moreover, listener's comprehension depends on other oral text: topic familiarity, discourse signals, word familiarity and conceptual difficulty of the text.

### **Background knowledge about the topic, text, structure, schema, and culture**

Student's background knowledge about a spoken text can have a profound impact on their ability to understand what has been said. Without a schema, understanding a text can be extremely difficult. Most people found it extremely difficult to recall the story exactly even after repeated readings; where the elements of the story failed to fit into the listener's schemata, they were omitted or changed into more familiar forms. This process shows that the degree to which information in the passage conforms to the listener's existing knowledge base determines how easy it is to understand.

### **Findings and Discussion**

After the data analysis, the results of six listening tests came out and they were compared for assessment. Among the three levels, word level was the easiest one for the students. Multiple choice test got 78% and it is considered as easy level. Gap-filling test got 58% and it is considered as difficult level. For phrase level, table completion got 47% and for gap-filling 41%. So, both are difficult levels. Sentence level is the most difficult level for the students. Answering the questions test got 40% and it is considered as difficult level. The score for dictation is 32% and it is difficult level. The following table shows different results and the difficulty level of tests can be found out easily.

<b>Level</b>	<b>Test Types</b>	<b>Difficulty level</b>	<b>Discrimination index</b>	<b>Correlation</b>	
<b>Word Level</b>	Multiple Choice	78%	0.31	- 0.94 (negative)	easy
	Gap-Filling	58%	0.33	- 0.53 (negative)	difficult
<b>Phrase Level</b>	Table Completion	47%	0.25	0.30 (positive)	difficult
	Gap-Filling	41%	0.30	- 0.06 (negligible)	difficult
<b>Sentence Level</b>	Answering the Questions	40%	0.18	- 0.70 (negative)	difficult
	Dictation	32%	0.28	0.33 (positive)	difficult

According to the results, only one test is easy level. The item difficulty index ranges from 0 to 100; the higher the value of item difficulty, the easier the question. The other tests are between 60% and 30% range. For discrimination index, D: 0.0-0.19 is defined as poor item, D: 0.2-0.29 is acceptable and D: 0.3-0.39 is good and D: >0.4 means excellent. Out of six tests, three types of test were found with discrimination index of 0.31, 0.33 and 0.30. Another discrimination index of two tests are acceptable. Only one test showed poor mean discrimination index D: 0.18. The item will have low discrimination if it is so difficult that almost everyone gets it wrong or so easy that almost everyone gets it right.

Negative correlation between difficulty index and discrimination index signified that with increase in difficulty index, there is decrease in discrimination index. In this paper, three types of test show negative relationship because their difficulty level increase and the differentiation is also high. One test is no or negligible relationship. Another two tests have positive relationship.

With reference to Harold S. Madsen (1983), the difficulty level of each test is identified. In these six tests, there is no *too difficult level* for second year English Specialization students. It is generally admitted that listening is a difficult skill for students of English as a second language. Although it is a complex and difficult skill, it can be practiced and mastered. The important issue is to find out what difficulties students encounter during listening practice. According to the needs analysis, students' needs and their wants vary individually. Students must be taught step by step through practice and group activities. So, they will get pleasure and knowledge through studying practice.

Students often complain about speaking rate of native speakers because speaking rate of native speakers is important for non-native speakers to grip the given message. Moreover, it is well-known that pronunciation affects students' comprehensibility. Only a few students believe that speakers' pronunciation in recordings does not affect their level of comprehension. The vast majority of the students admit that speakers' pronunciation in listening materials affects their listening comprehension. This implies that students need exposure to various English dialects.

When the students listen to unfamiliar speech, they hear an almost continuous chain of sounds. Inexperienced students do not actually hear the boundaries of words. Experienced students are able to break down this chain into separate words in their heads because they are familiar with the sounds and can create meaningful words with them. Errors at this level may impede the students in the correct understanding of the spoken utterance. Intonation is interested with pronunciation and it is known as the ability to vary the pitch and tune of speech. Stressing words and phrases correctly is vital to give the important parts of messages.

Next point is that knowledge of vocabulary is of uppermost important in understanding listening texts. Most of the students found it difficult to recall general vocabulary. Thus, the teachers need to pay more attention to revising lexis; otherwise, students find it difficult to activate learnt vocabulary.

Another problem is that a number of different types of knowledge are involved: both linguistic knowledge and non-linguistic knowledge. Linguistic knowledge is of different types, but among the most important are phonology, lexis, syntax, semantic and discourse structure. Then non-linguistic knowledge is knowledge about the topic, about the context and general knowledge about the world and how it works. This world knowledge can influence comprehension in several ways.

The major problem that occurs in listening is that most of the students think they can answer the tests if they catch all the words in listening texts. There is no need to hear every single sound in every single word to know. The thing they need to have is conceptual skills. So the teachers should train them to develop their conceptual skills and thinking.

To sum up, the first problem for the students is speaking rate. If the speaking is very fast, they will give up easily. Secondly, pronunciation and intonation in listening recording affects comprehension of many students. Thirdly, even pre-taught vocabulary might present difficulty in understanding. Finally, during listening activities, the students find it difficult to get specific information and want to hear each word. And the solutions for the first and second problems are that students need to be exposure to various speaking rate with different accents. The teachers encourage them to listen to spoken tests including songs. For the third problem, they need to expand their vocabulary by listening to different sources of materials. If they have enough vocabulary, they can catch the messages more easily and can do well in their tests. For the last one, the teachers teach them how to grab the specific or intended information of the speakers.

## Conclusion

The main purpose of this paper is to find out difficulty levels, discrimination indexes and their correlation. According to Harold S. Madsen, five types of test in this paper are defined as *difficult level* because all these tests are between 30% to 60% range. One type of test is defined as *easy level* as its result is ranged between 60% and 90%. There are no *too difficult level* and also no *too easy level*. In addition, discrimination index of each test can be identified by analyzing tests outcomes. Discrimination index of three types of test can differentiate well between weak and good students. Two are acceptable but they cannot discriminate distinctly. The last one is found as poor discrimination index. There is no excellent discrimination index in this paper. A very high difficulty level fails to discriminate and generally show a poor discrimination index. When difficulty level was analyzed along with discrimination index, correlation between them came out. Three types of test have negative correlation because one test is easy and two types of test are difficult. The result of one test is negligible as there is a slightly difference between difficulty level and discrimination index. Another two tests have positive correlation.

As the role of listening has become significant in English, students should give full attention. The students can improve their listening skills by doing a lot of exercise. In order to understand the nature of listening processes, they need to have knowledge of the language and the ability to apply that knowledge. However, it will take some time to be skillful in them. So, the teachers should train the students with effective strategies and methods step by step and encourage. These strategies seek to involve the students actively in the process of listening. And the students should have exposure to spoken English outside the classroom.

## Implication

In public school and universities, teachers or testers should look carefully at their test items to see if the majority of the questions that demand higher levels of thinking skills are too difficult or do not discriminate well. And if the majority of the questions that demand lower levels of thinking skills, especially recall questions, are too easy or do not discriminate well among students. This leads the way for rote-learning and students behave like parrots. Their critical thinking and problem-solving, adaptability, initiative, effective oral and written communication, synthesizing and analyzing information, and curiosity and imagination skills will be retarded and they will not pay any attention to lessons at their schools and universities. They will rely on private schools since they assume that government education sector will not give any guarantee for their future prospects. The paper wants to encourage all teachers to develop tests which are not too easy and difficult and to notice what is happening around them by comparing with international standard. What we all teachers should do is that we need to advance the quality of teaching and learning by maintaining high and rigorous standards for what accomplished teachers should know and be able to do and advocating related education reforms to cope with globalization while marching towards democratic nation.

## Acknowledgements

I would like to give thanks to Dr. Hla Htay, Rector of Dagon University who encouraged everyone to do the research at every department of Dagon University. I would like to express my heartiest thanks to Daw Swe Swe Pwint, Professor, Head of English Department, Dagon University, for her permission to write this research and for her encouragement during the writing of this research. I would like to acknowledge Dr. Thi Thi Tun, Professor of English Department, Dagon University, for her patient guidance and invaluable advice to complete this paper. I would like to extend my deepest thanks to Daw Hnin Hnin Aung, Associate Professor of English Department, Dagon University, for her constant encouragement in the preparation of this research. A special note of thanks goes to my friends who gave me sound advice and help in writing this research.

### References

- Bloomfield, A., (2010), "What Makes Language Difficult?", (unpublished), University of Maryland, Maryland
- Brown, J. D., (1996), "Testing in Language Programs", Prentice Hall Regents, Prentice-Hall, Inc., New Jersey
- Graham, A., (1999), "Teach yourself: Statistics", Hodder Headline Plc., London
- Harmer, J., (2005), "The Practice of English Language Teaching", Pearson Education Limited, Edinburgh
- Honby, A.S., (2005), "Oxford Advanced Learner's Dictionary of Current English", 7<sup>th</sup> Ed., Oxford University Press, New York
- Heaton, J.B., (1988), "Writing English Language Tests", Longman Group UK. Ltd., Longman
- Hughes, A., (2003), "Testing for Language Teachers", Cambridge University Press, Cambridge
- Madsen, H.S., (1983), "Techniques in Testing", Oxford University Press, Oxford
- Richards, J.C., (1998), "New Interchange 3", Cambridge University Press, Cambridge
- Valette, R.M., (1977), "Modern Language Testing", Harcourt, London

### Online Resources

English for Specific Purposes World, Online Journal for Teachers. (ESP World) at: [http:// esp-world.info](http://esp-world.info)