DEVELOPING LISTENING COMPREHENSION TEST FOR STUDENTS OF ENGLISH EDUCATION DEPARTMENT

Oleh : Budi Setyono

FKIP Universitas Negeri Jember Jalan Kalimanatan 37 Jember
Email : bssetyono@gmail.com
Jalan Raung 3/K13Jember

Abstract.

Memiliki kemampuan mengembangkan tes sebagai salah satu instrumen asesmen merupakan salah satu kompetensi yang harus dimiliki oleh seorang pendidik. Tes merupakan instrumen yang banyak digunakan oleh pendidik untuk mengumpulkan data mengenai belajar peserta didik. Dengan memberikan tes, pendidik dapat memonitor kemajuan belajar, mengukur hasil belajar, serta mengevaluasi efektifitas pembelajaran. Agar dapat mengembangkan tes kemampuan listening yang baik, dosen pengampu matakuliah Listening dipersyaratkan memiliki pemahaman yang mendalam terhadap ciri-ciri tes yang baik serta langkah-langkah dalam mengembangkan tes. Untuk tujuan tersebut, artikel ini akan membahas bagaimana mengembangkan tes kemampuan listening yang valid dan reliabel di jurusan pendidikan bahasa Inggris. Bagian pertama akan mengupas pengertian kemampuan listening, kemampuan listening yang dapat diukur, serta beberapa format tes kemampuan listening. Pada bagian berikutnya akan dibahas konsep tes yang baik dilihat dari aspek validitas, reliabilitas dan kepraktisan. Bagian akhir dari pembahasan ditutup dengan prosedur dalam mengembangkan tes kemampuan listening di jurusan pendidikan bahasa Inggris.

Kata kunci: test development, listening comprehension test, English education

INTRODUCTION

Those choosing profession as teachers are required to possess a set of competencies in order to grow professionally as long as their teaching careers. In general, teachers are required to have competencies in designing instructional plan, implementing instructional plan, assessing instructional process and product, and providing guidance for their students. With such competencies, teachers are expected to be able to present satisfactory instructional practices which in turn can develop students' knowledge, skills, and attitudes.

Assessing process and product of instruction is an important thing to be done by the teacher. It aims to monitor students' progress as learners, the effects of instruction on the students, and the effectiveness of instruction. Teachers can use several kinds of techniques in order to assess both instructional process and product. Observation, interview,

Budi Setyono adalah dosen FKIP Universitas Jember

questionnaire, and test are some examples of assessment techniques.

For classroom use, test as an instrument of assesment is widely used by the teachers. In this context, test is used for the purpose of measuring students' mastery of the materials of a course/ lesson having been taught. Such kind of test is known as achievement test as it is used to know to what extent students have already mastered the teaching materials. In terms of the time for conducting the test, achievement test can be classified into formative and summative test. Formative test or progress test might be given either at the end of each chapter/unit or several chapters/units, middle of instructional program, while summative test is given at the end of instructional program. On account of its relation to the instructional process, valid achievment test must be constructed on the basis of objectives and instructional contents.

As an instrument of assessment, the results of achievement test cannot only function to assign students grades, but also function to give feedback for the improvement of instructional process. From the analysis of test results, teachers can identify knowledge, skills or objectives having been mastered by their students. On the basis of the test results, teachers are expected to modify their instructional strategies, such as modifying the instructional materials and/or modifying methods of instruction.

Considering the importance of test, the achievement test must be well-constructed and the test results – i.e. students' scores – must be analyzed in oder to improve its quality for future use. Despite the importance of test, most teachers rarely pay serious attention to the development of test and do analysis on its results. To develop a good test, the test constructor is required to know its characteristics as well as procedures for developing it. Classroom teachers who are also test constructors should also understand the principles of test development in order to be able to develop a good test.

DISCUSSION

Listening Comprehension

Before constructing a test of listening comprehension, it is necessary for the test writers to understand the concept of listening comprehension. By understanding this concept, the test writers are expected to be able to identify the types of listening abilities intended to be measured from the test takers. Inability to identify the types of listening abilities to be measured will affect the quality of the test they produce, as the test writers cannot show the kinds of listening abilities they want to measure in reference to the theory underlying the items of test. As a result, their tests cannot be categorized as good tests, since the tests lack of contruct validity.

In language teaching, listening and reading are grouped as the receptive skills, whereas speaking and writing are grouped as the productive skills. In the history of English Language Teaching, listening as a receptive skill is frequently neglected. In the era of audiolingualism, for example, the focus of language instruction was on oral production rather than on comprehension. Listening is used as a means to teach oral production. During this era listening is applied

only to the auditory processing of a very short speech segment, i.e. word, phrase, sentence in order to reproduce it (listening to repeat). Thus, the objectives of Listening comprehension were the mastery of oral grammar and pronunciation rather than listening for meaning. Due to this Newmark and Diller (in Morley, 1983) suggested to give emphasis on the development of Listening comprehension skill rather than as a foundation for speaking. The different concept of understanding listening as a means to teach other skills and listening as skill of its own that must be developed affect how language teachers teach and test Listening comprehension.

According to Morley (1983), listening comprehension is defined as listening to understand meaning. In addition to this, Rost (2001) states that listening is used to refer to a complex process that allows us to understand the spoken language. More elaborately, Buch (1990) states that the listening comprehension is a massively parallel process of taking advantage of information from a large number of sources, both linguistic and nonlinguistic and it may not be possible to separate out individual variables.

These definitions give indication that listening is a process of understanding meaning from the spoken language, and in an attempt to understand the messages linguistic and non linguistic knowledge will play a major role.

In the process of understanding meaning from the spoken language, there are two kinds of distinct process involved, i.e. bottom-up processing and top-down processing. Bottom-up processing refers to a process of decoding a message that the listener hears through the analysis of sounds, words, and grammar, while top-down processing refers to using background knowledge to comprehend a message (Nunan, 1993). Both processes are assumed to take place at various levels of cognitive organization: phonological, grammatical, lexical and propositional.

Components of Listening Comprehension

From the processes involved in understanding meanings of the spoken language, it can be identified that in listening comprehension, linguistic and nonlinguistic components equally play a major role.

Listening comprehension includes the recognition of words, structures, and pronunciation features. Subskill listening is concerned with the linguistic componenets of language, and macro listening deals with the broader communication, which concernes with the exchange of facts and ideas, as well as interpreting the speaker's intentions.

In the summary checklist of operations for testing Listening Comprhension proposed by Weir (1993), it can be specified what kinds of listening abilities that the test writers want to measure.

Basically, the summary checklist (see Table 1) classified four categories of comprehension to be measured in listening comprehension test: i.e. (a) direct meaning comprehension, (b) inferred meaning comprehension, (c) contributing meaning comprehension (microlinguistic), and (d) listening and writing (note taking from lecture, telephone conversations, etc.).

Table 1:
Summary Checklist of Operations (Listening Comprehension) taken from Weir, 1993.

(a)	*Direct meaning comprehension* Listening for gist Listening for main idea (s) or important information; includes tracing the development of an argument, distinguishing the main idea (s) from supporting detail, differentiating statement from example, differentiating a proposition from its argument, distinguishing fact from opinion when clearly marked Listening for specifics; involves recall of important details Determining speaker's attitude/ intentions toward listener/topic (persuassion/explanation0 where obvious from the context)
(b)	*Inferred meaning comprehension* Making inferences and deductions; evaluating content in terms of information clearly available from the text Relating utterances to the social and situational context in which they are made Recognizing the communicative function of utterances Deducing meaning of unfamiliar lexical items from the text
(c)	*Contributory meaning comprehension (microlinguistic):* Understanding phonological features (stress, intonation, etc). Understanding concepts (grammatical notions) such as comparison, cause, result, degree, purpose. Understanding discourse markers Understanding syntactic structure of the sentence and clause, e.g. elements of clause structure, noun and verb modification, negation. Undersatnding lexical cohesion through lexical set membership and collocation Understanding lexis.
(d)	*Listening and writing (note taking from lecture, telephone conversation, etc.)* Ability to extract salient points to summarize the whole text, reducing what is heard to an outline of the main points and important detail. Ability to extract selectively relevant key points from a text on a specific idea or topic, especially involving the coordination of related information

Testing Listening Comprehension

In relation to language teaching, listening is an important skill that cannot be neglected. The development of this skill must always be monitored and measured through listening test. Ability to understand meanings of the oral language becomes the target of listening comprehension test (Djiwandono, 1996).

According to Madsen (1983), test of listening skill is divided into two categories. The first category of test is listening as a tool to evaluate something else (*sub-skill*). In this matter, listening is used to measure word mastery, proficiency in grammar and pronunciation, and more advanced integrative skill. The second category of test is using listening to evaluate the listening comprehension skill. In summary, listening sub-skill test focuses on the liguistic components of language, while the listening comprehension test is concerned with the broader communication, i.e. it is not concerned with pieces of language but with the exchange of facts and ideas as well as interpreting the speaker's intetions.

Above all, Heaton (1990) argued that 'the ability to hear sound differences is not necessarily the same as the abililty to understand spoken messages'. Following Heaton's, Weir (1993) also argued an ability to discriminate

phonemes does not imply a capacity to comprehend verbal messages. This statement indicates that sub-skill test of listening cannot be categorized as test of Listening comprehension.

Currently test of Listening comprehension is concerned with testing the communication of meaning instead of testing the structural understanding. The communication of meaning refers to the exchange of of messages (i.e. the sending and receiving of information) from speaker(s) to listener(s). So, testing the communication of meaning means testing students' ability to extract or understand information from contexts rather than the pieces of information.

Text Types Used in Listening Comprehension Test

Text is the written record of a communicative event which conveys a complete message (Nunan, 1993). Texts may vary from single words to books running to hundreds of pages. As used in instructional materials, texts selected for testing Listening comprehension can be taken from two broad categories of oral language, i.e. monologue and dialogue (Nunan (1991) cited in Brown, 2001). Monologues include planned monologues, such as speeches, newsbroadcast, and other prewritten material and unplanned monologues, such as impromptu lectures and long stories. Dialogues can be exchanges promoting social relationships (*interpersonal*) and those having purpose to convey propositional or factual information (*transactional*.

For the types of texts, the important thing to consider for the purpose of testing listening comprehension is the level of difficulty. It must be attempted that the texts must be within the students' instructional level. This means that the texts should not be too difficult or too easy for students to digest. The difficulty level of the texts can be seen from the concept load, complexity of the syntax, and the breadth of vocabulary in the texts.

Besides, another important thing to consider is the length of texts and speed used by the speakers. In order to choose the texts properly, the texts must be adjusted with the students' level. Short texts might be suitable for elementary-level students, while texts with normal and faster speed might be suitable for more advanced-level students.

In relation to the text contents, topics should be attempted to cover the range of situations which represent samples of language use both in monologues and dialogues. The possible topics may be related to everyday situations, such as daily routines, holidays, directions, flight announcements, telephone messages, reservation of hotel accomodation, and description (jobs/locations of objects/ peoples' physical characteistics). Other possible topics may be related to more formal situations, such as interviews, short lectures/talks on a variety of subjects, news, documentaries, and sports commentaries.

Texts selected for listening comprehension test have to contain components of listening comprehension abilities intended to be measured. This means that in reference to the texts test developers can formulate questions for measuring listening abilities of the test takers. The categories of comprehension, as stated previously, are categorized into direct meaning comprehencion

(microlinguistic), and listening and writing (note taking from lecture, telephone conversations, etc).

Possible Test Formats for Listening Comprehension Test

The main objective of listening comprehension test is to evaluate test takers' comprehension of the text. To evaluate test takers' listening comprehension abilities, test developers may use formats that are suitable to elicit the particular listening abilities intended to be measured. As different test formats would measure different aspects of language ability, it is required that test developers understand the characteristics, uses, advantages, limitations, and rules for construction of the test formats (Gronlund, 1985). Several possible formats that can be employed to test Listening comprehension, i.e. multiple-choice, short-answer questions, completion, true-false, matching, information transfer, dictation, and listening recall will be discussed in the following.

(1) Multiple-Choice

Multiple choice items take many forms but the basic structure is: there is a stem, and a number of options, one of which is correct, the others being distractors (Hughes, 1996). There are several advantages of using multiple choice formats. First, scoring of multiple choice test is perfectly reliable. Second, the multiple-choice items can be used to measure almost any measurable mental process (Schoer, 1972). Lastly, it is possible to include more items using multiple-choice test. Apart from its advantages, multiple choice tests have some problems, such as chances for guessing, difficulties in writing good items, harmful backwash, and cheating facilitacion.

(2) Short Answer

Short-answer item is the supply-type test items that can be answered by a word, phrase, number, or symbol. The short-answer test item uses a direct question. It is suitable for measuring a wide variety of relatively simple outcomes, chiefly to measure the recall of memorized information. This format is advantageous as it is easy to construct and reduces guessing. The limitations that restrict the use of short answer item are unsuitability to measure complex learning outcomes and the difficulty of scoring, except if the items need only one correct response.

(3) Completion

The completion item is also a supply-type test items that can be answered by a word, phrase, number, or symbol. The completion item consists of an incompleten statement. Like short answer item, the completion test item is also suitable for measuring a wide variety of simple outcomes, chiefly to measure the recall of memorized information. This format is advantageous as it is easy to construct and reduces guessing. The limititations restricting the use of completion item are unsuitability to measure complex learning outcomes and the difficulty of scoring, except if the items need only one correct response.

(3) True-False

The alternative-response test item consists of a declarative statement that the student is asked to mark true or false, right or wrong, correct or incorrect, yes or no, fact or opinion, and agree or

disagree. Because the true-false option is the most common, this item type is frequently referred to as the true-false test item.

The most common use of this item is in measuring the ability to identify the correctness of statements of fact, definitions of terms, statements of principles, and the like. A common criticism of this format is that student may be able to recognize a false statement as incorrect but still not know what is correct. Two advantages, although not very valid, are ease of construction and a wide sampling of course material can be obtained. The serious shortcoming is that it measures the learning outcomes in the knowledge area. Another limitation is its susceptibility to guessing. With two alternatives, a pupil has a fifty-fifty chance of selecting correct answer on the basis of chance alone.

(5) Matching

Matching items consist of two paralled columns, with each word, number, or symbol in one column. The items in the column for which a match is sought are called premises, and the items in the column from which the selection is made are called responses. The pupils' task is to identify the pairs of items that are to be associated on the basis indicated. The typical matching format is used to measure factual information based on simple associations. The pupil's task is to relate two things that have some logical basis for association. The advantages of matiching items are: (1) the possibility to measure a large amount of related factual material in a relatively short time, (2) ease of construction. As with tue-false item, poor items can be rapidly to construct, but good items require a high degree of skill. The limitations of matching items are: (1) it is restricted to measure factual information based on rote learning, and (2) it is susceptible to the presence of irrelevant clues.

(6) Information Transfer

In information transfer, information which is transmitted verbally is transferred into nonverbal forms, such as labeling or drawing diagrams or picture, recording routes or locating buildings on a map, completing tables or numbering a sequence of events. Using this format students' range of skills can be tested. It is efficient to test an understanding of sequence, process description, relationships in a text and classification. However, this format is not suitable to test the skills of inferred meaning comprehension or determining the speakers' attitudes. The problems with this format are: (1) it is difficult to find spoken texts which fit into this format, (2) it is often difficult to create drawings or produce illustrations.

(7) Dictation

In dictation, test takers will listen to the dictated texts, e.g. references or scientific laws. The test takers' task is to write everything down. Dictation tasks may involve recall of details at the level of direct meaning comprehension and contributory meaning comprehensions, e.g. discriiating phonological units and determining word boundaries. These tests are are easy to construct, easy and quick to administer. With adequate training these tests are also easy and quick to mark. As indirect test, the main problem is concerned with how students' performance in the test can be translated into a direct statement of proficiency.

(8) Listening Recall

Using this format, test takers are asked to complete the deleted words from a multilated (i.e. passage from which some words have been deleted). The words deleted are normally content words felt to be important to understanding of discourse. Asuggested procedure to administer the test is as follows. First, students listen to the complete text and take notes. Then, students are given the mutilated passage. Lastly, students listen to the complete text and fill in the blanks while listening. An advantage of the test is that it is easy to construct, administer, and mark. Decisions must be taken in advance as to whether answers will be marked by an 'exact word' or an 'acceptable alternative'. The major drawback of this format for the tester is the difficulty in saying with what is being tested. Where only one word is deleted, it may not be testing anything more than an ability to match sounds with symbols, aided by an ability to read the printed passage containing the gaps.

Features of a Good Test

In order to develop a good test, test writers are required to think about its validity, reliability and practicality. These three caharacteristics must be fulfilled in developing any kind of test.

Validity

Thinking about the validity of a test may initially involve defining what it is that the testers wanted to measure. If they cannot define what they wanted to assess, they cannot determine the degree to which the test is measuring it. Brown (1996) defines validity as the degree to which a test measures what it claims, or purports, to be measuring. In other words, the constructed test items must really measure the intended ability or skill to be measured. Test validity can be established in three ways: content validity, construct validity, and criterion-related validity.

Content validity means that the test items constructed must reflect the contents of a particular course. In order to investigate content validity, testers must decide wheher the test is a representative sample of the content of whatever the tes was designed to measure (Brown, 1996). Thus, the goal of content validation should always be to establish an argument that the test is a representative of sample of the content that the test claims to measure. In relation to the classroom test, content validity can be established by way of logical validation in the form of qualitative information about the objectives and the contents of particular course. On the basis of the course objectives and contents, testers can construct a table of specifications.

Constuct is defined as an attribute, proficiency, ability or skill in psychological theories (Brown, 1996). It exists inside the head, very often it is unobservable, and thus it is difficult to measure . Some examples of construct in relation to the topic of language testing among them are speaking ability, writing ability, Listening comprehension ability, and reading comprehension ability. In order to test those abilities, testers must be able to define those constructs. For example, ability in Listening comprehension is defined as testee's ability to comprehend messages from a test being heard. The comprehension itself can be divided into literal, infeential and evaluative comprehension.

Having identified those concepts, then the testers can write test items that will measure the intended concept. Construct validity can be established by way of logical and empirical validation (Djiwandono, 1996). Logical validation can be gained through reasoning in the form of qualitative information about kinds of ability to be measured in the test, while empirical validation can be established by way of correlating the scores gained from the developed test and other (established) test in the same field having similar construct.

Criterion-related validity refers to an examination on the scores of the new test that the testers are developing and the scores of other test that is already a well-established measure of the construct involved. It can be established through empirical validation. By employing criterion-related validity, one group of students will take two tests: the developed test and other test supposed to be a well-established test of the construct under investigation. If the two tests are administered to the same group of students at different times, it is termed as predictive validity. The scores obtained by the same group of students from the two tests will be correlated to calculate the correlation coefficient or a validity coefficient.

Reliability

Reliability is a concept referring to the consistency of the results of the test. An observed score on any test is a composite of two components, i.e. a true and an error component (X=T+E). When a student takes a test many times, the observed score would be slightly different each time. The observed score is only an estimate of the true score because fluctuations in environement (changes in test administration) and psychological changes (tiredness) could change the observed score (Fulcher & Davidson, 2003). Reliability can be established internally or externally indicated by statistical figure known as correlation coefficient index, ranging from 0.00 to 1.00 (Weir, 1990). Internal reliability is measuring the internal consistency of the test by way of giving test once. In practice, most tests report measures of internal consistency as reliability coefficients. These measures are simply a measure of mean inter-item correlation, or how well items correlate with each other. Internal reliability coefficients are affected by the following factors. The first factor is the number of items. Increasing the number of test items will increase reliability. The second one is variations in item difficulty. Items having equal difficulty will increase reliability, while items having a range of facility values will decrease reliability. The next factor relates to dispersion of scores. If the test scores are homogeneous or no spread of scores, reliability will decrease. Lastly, it is related to level of item difficulty. Items with facility values of 0.5 maximize item variance, and so increase test reliability (Fulcher & Davidson, 2003).

The internal reliability can be obtained either through split-half method (by applying Spearman-Brown formula) or Kuder-Richardson method (by applying Kuder-Richardson formula K-R21 or K-R20). External reliability is measuring the external consistency of the test that can be obtained using test-restest method and equivalent-forms method. To estimate reliability using the test-retest method, the same test is

administered twice to the same group of test takers in two different time periods. The resulting test scores are correlated, and this correlation coefficient provides a measure of stability indicating how stable the test results are over the given period of time.

In estimating reliability using the equivalent-forms method, two equivalent tests o parallel tests are administered to the same group of test takers in close succession. The equivalence of the two tests are attempted to cover the aspects of ability to be measured, test formats, time duration to develop the tests, contents coverage, level of difficulty, and the number of test items. The resulting test scores are correlated, and the correlation coefficient provides a measure of equivalence indicating the degree to which both test are measuring the same aspects of ability.

Practicality

Practicality, another important characteristic of a good test, deals with something which is practical instead of theoretical, such as ease of administration, ease of scoring, and ease in interpreting the test results (Djiwandono, 1996;Gronlund, 1985). In terms of administration, practicality of test refers to the provision and use of facilities which are not special and complicated. For example, a practical listening test can be administered in a classroom using a portable tape-recorder. In terms of scoring, practicality of test concerns with the ease of examining and scoring of the test and minimum amount of time required to score the test. Certainly, this practicality should not sacrifice the scoring accuracy. Efficient scoring can be established by providing simple scoring directions, simple scoring keys, separate answer sheets, and if possible the scoring machine (Gronlund, 1985).

Procedures in Developing Listening Comprehension Test

To give the real context of test development, the test development procedures in the following discussions are applied in designing the listening comprehension test at English Education Department.

Examining Objectives and Listening Materials through Syllabus

As an achievement test, the listening comprehension test is constructed based on course objectives and contents of listening comprehension. The test relates to the instruction and measures students' learning outcomes of listening comprehension materials having been taught. Based on the examination of the syllabus developed by a lecturer of Listening II Course at the Faculty of Teacher Training and Education, the University of Jember, the listening comprehension test has an objective of 'developing students' ability to understand spoken English at the itermediate level aiming at literal and inferential comprehension of short statements, dialogues, narrative, descriptive, and expositional types of texts'. On the basis of this objective, the texts for use as listening comprhension materials are short statement, dialogue, narrative, descriptive and expositional texts. Dialogue texts are taken from listening course books entitled 'Taks Listening' by Lesly Blundell and Jackie Stokes, 'Learning to Listen' by Alan Maley and Sandra Moulding, and Building Skills

for the TOEFL' by Carol King and Nancy Stanley. Narrative, descriptive and expositional texts are taken from listening course books entitled 'Listening Comprehending' by M.H. Combe Martin and some texts are taken from 'Building Skills for the TOEFL' by Carol King and nancy Stanley. (Syllabus of Listening Comprehension at English Education Department of FKIP, the University of Jember).

Stipulating the Intended Ability to be Measured in the Listening Test

Referring to the course objectives and contents of Listening Comprhension, the Listening comprehension test is developed to measure the listening comprehension ability of the thrid semester students taking Listening comprehension II course at the English Department of FKIP, Jember University.

The intended listening ability to be measured in the test is limited on the literal and inferential comprehension or direct meaning comprhension and inferential meaning comprehension (Weir, 1996). In these categories of comprehension, students are expected to be able to comprehend information stated explicitly and implicitly in the listening texts.

Drawing Table of Test Specification and the Test Formats

By considering coverage of materials to be tested, th alloted time, the number of students in two parallel classes, ease and reliability of scoring, the listening comprhension test is developed using objective test. The formats selected are multiple-choice and short-answer question. multiple choice format is suitable to test broad coverage of test materials, whereas short-answer question format is chosen to minimize chance of guessing. Entirely, there are 40 test items; 35 items are formatted in multiple choice and 5 items are formatted in short answer. The test consists of six parts (Part A, B, C, D, E, and F). Part E to E (no. 1 to 35) are developed using multiple choice format, while Part F (no. 36 to 40) is developed using short answer format. There are 23 items intended to test the ability to comprehend explicitly-stated information and 19 items intended to measure ability to draw inferences from texts.

What is tested in Listening Comprhension and its proportions can be seen from the table of specification (see Table 2). Without a table of specification, it is very difficult to check the relevance of test items with the objectives of instruction.

Table 2.
The Table of Test Specification

Format	Number of items in the test	Test Objectives		Total
		Comprehend explicit messages	Draw inferences	
multiple choice	1 to 35	22 items (55%)	13 items (32.5%)	35 items (87.5%)
Short-answer	36 to 40	4 items (10%)	1 item (2.5%)	5 items (12.5%)
Total	26 items (65%)	14 items (35%)	40 items (100%)	

The time allocated to do the multiple choice test is 45 minutes, while the time alloted to do short-answer test is 15 minutes. Entirely, it takes 15 minutes to play the cassete for 5 types of texts in multiple choice test, and the rest of time 27 minutes is spent to do the test and check the answers. In addition to this, the time needed to play the cassette for 1 type of text (played twice) in short answer test is 3 minutes, and 12 minutes is allocated to do the test and checkthe answers.

The directions of the test consist of general instructions and specific directions of each part. General directions direct students to answer the test on the answer sheet given, to pay attention to the number of times the cassette played in each Part, and to answer the questions while listening to the cassette. Specific directions in multiple choice test direct the students to listen carefully, to choose the correct answers. Specific directions in the short-answer test direct students to listen to the text carefully, to answer the questions briefly, and to check the answers.

Selecting Texts for the Listening Comprehension Test

Texts for the listening comprhension test are selected by considering the difficulty level, length and speed, contents, and the availability of listening comprehension components in the texts. To meet these requirements, all texts used in this test were the texts that were taken from course books of listening used in the listening comprehension course. The types of texts used are short statement, dialogue, description-narration, and exposition. Ten short statements, two texts of dialogues and one text of exposition were taken from TOEFL preparation books, while one dialogue and one descriptive-narrative text were taken from course books of listening comprhension.

Texts taken from various sources are strictly selected in order to meet the students' instructional level. Seen from the concepts in the texts, all texts taken from TOEFL preparation books are not too difficult to understand as the topics o dialogue texts are related to everyday situations, i.e. asking for directions and seeking for part-time job on campus, and one topic is about protecting environment. From the aspect of syntax, the expressions used to exchange ideas in dialogues do not indicate the complex construction of sentences. Instead, they are expressed in simple constructions (simple questions and statements) which are commonly used in asking for direction and part time job. An example is Can you tell me how to get to the Music Building from here ? Lastly, from the aspect of vocabulary, the words used to express meanings in the texts are mostly not difficult to understand. For some difficult words, such as rehearsal and boasts, students are expected to guess from contexts.

In addition to this, all texts selected are delivered in normal rates of speed on topics and situations which represent samples of language use both in monologues and dialogues. The texts representing the topics of everyday situations are Phoning a Flat Owner, Asking for Direction to the Music Building, Finding Part-time Job, and A Thief in the Bus. Only one topic is related to an academic situation, i.e. a topic about environmental protection as found in the text entitled Sierra Club.

In connesction with components of listening comprehension in the texts, the texts selected can be used to measure students' understanding of the texts at the level of direct meaning comprehension and inferred meaning comprehension.

Writing Test Items

The listening comprehension test items are developed using multiple choice and short-answer formats. Each multiple choice item in the Listening comprehension test consists of a stem and four options containing three distracters and one correct answer. The stem is in the form of a question or incomplete statement. The multiple choice items in the test are not written, rather, some items are directly taken from the original source, and some others are the modified items. Test items in part A, Part D, and Part E, for example, are directly taken from excercises in a workbook prepared for TOEFL preparation test entitled Listening to TOEFL and Longman Preparation Course for the TOEFL. Some items in these parts are intended to test comprehension both at the literal level, and some others are intended to test comprehension at the inferential level. Other multiple choice items in Part B and Part C are not directly taken, rather they are modified from items of test originally written in short-answer format. On account of it, options must be provided in developing multiple choice items in Part B and C. Besides, some questions are also reformulated to make them clearer.

Test items in part F are also modified items as oroginally Part F are written in multiple choice format. On account of it, some stems in the original item are reformulated and all options are eliminated. The following example is a modified item by way of modifying the question of the original item and leaving out the four options.

Original item : Which of the following is not listed in the passage as a publication of the Sierra Club ?
a. A newsletter
b. A magazine
c. Statistical studies
d. Books

Modified item : What is one of the publications of the Sierra Club ?

The students have to provide short answer. In this case the answer is only one of the Sierra Club's publications, that is a newsletter or a magazine or books. Students answering two or three publications will be considered wrong as the intnded answer is one of the publications. As the answer is stated explicitly in the text, the item is measuring literal comprehension .

Refining Test Quality

In order to refine the quality of listening comprehension test, the two refining procedures are conducted, i.e. asking listening comprehension instructors to give feedback via questionnaire and piloting the test. In the questionnaire, they are asked to give comments on the test instruction, the alloted time to do test, the number of items, the quality of recordings, the coverage of test materials, and the difficulty level of test. Next, the listening comprehension test will be piloted to the students of the English Department joining Listening course in different

class. The quality of test items will be identified through item analysis; whereas the reliability coefficient index will be computed through internal consistency by applying KR-21 formula.

Analyzing Test Items

The process of inspecting individual test items is familiarly known as item analysis. More formally, item analysis is defines as the systematic evaluation of the effectiveness of the individual items on a test (Brown, 1996). The following will describe three ways that can be applied the in conducting item analysis, i.e. item facility, item discrimination, and distracter effciency analysis.

Item facility (IF) or item difficulty is a statistical index used to examine the percentage of students who correctly answer a given item. To calculate the IF of Listening comprehension, add up the number of students who correctly answered a particular item, and divide that sum by the total number of students who took the test. The formula looks like the following:

$$IF = \frac{N\ correct}{N\ total}$$

where Ncorrect = number of students answering correctly
Ntotal = number of students taking the test

(Brown, 1996)

The result of this formula is an item facility value that can range from 0.00 to 1.00 for different items. An item having IF .27, for example, indicates that 27% of students answered the item correctly. It is categorized as a very difficult question because more students (73%) are unable to answer the item. In contrast, an IF of .96 indicates that 96% of the students answered correctly. This item is categorized as a very easy item because almost everyone responded correctly. The accepted item would be IF having range between .20 - .80. Item discrimination (ID) indicates the degree to which an item separates the students who performed well from those who performed poorly. These two groups refer to the upper and lower third, or 33%. Some test developers will use the upper and lower 27%, and some others will use 25% to ca Listening comprehensionulate ID.

There are several steps to follow to calculate Listening comprehensionulate ID value. First, divide raw scores into three groups (top group, middle group and bootom group). Apply either 33% or 27%, or 25% to determine the upper and lower groups. Once the data are sorted into groups of students ID index can be ca Listening comprehensionulated. To do this, calculate Listening comprehension the IF for the upper and lower groups separately for each item. This is done by dividing the number of students who answered correctly in the upper group by the total number of students who answered correctly in the lower group; then divide the number of students who answered correctly in the lower group by the total number of students in the lower group. Finally, to calculate Listening comprehension the ID index , the IF for the lower group is substracted from the IF for the upper group on each item as follows:

ID = Ifupper – Iflower
Where ID = item discrimination from an individual item
IFupper = item facility for the upper group on the whole test
Iflower = item facility for the lower group on the whole test

(Brown, 1996)

The result of this formula is an item discrimination index (ID) that can range from 1.00 to -1.00. Naturally, ID indexes can take on all the values between +1.00 and – 1.00 as well. An item discrimination index of 1.00 is very good because it indicates the maximum contrast between the upper and lower groups of students, i.e. all the high-scoring students answered correctly, and all the lower-scoring students answered correctly.

Based on the ID index, some items can be categorized as good, and some others are categorized as poor items and need improvement. The following is a guideline that can be used for making decisions based on ID:

.50 and up	: good
.20 to .50	: fair
Below .20	: poor
0	: no discrimination
-(negative)	(negative)

Items having ID below .20 still need improvement, whereas items having no discrimination or negative index must be eliminated.

Distracters, options that are counted incorect, should divert or pull away test takers from the correct answer if they do not know which answer is correct. In other words, it can be said that good distracters must be able to divert the test takers who do not know the answer. Therefore, in order to provide good distracters, distracter efficiency analysis must be done. According to Brown (1996) distracter efficiency analysis has a main purpose of examining the degree to which the distracters are attracting the test takers who do not know the correct answers. In order to do this for an item, the percentages of the test takers who chose each option is analyzed. If this analysis can give the percentages choosing each option in the upper and lower groups, the information is useful. This means that the upper group should have more frequency of correct answers compared to the lower group.

CONCLUSIONS

In developing listening comprehension test, test developers need to possess a set of knowledge related to some aspects of test development. The most important aspect is having good knoweldge about the construct of listening comprehension. By understanding this concept, the test writers are expected to be able to identify the types of listening abilities intended to be measured from the test takers. Conceptually, listening comprehension test is concerned with the exchange of facts and ideas as well as the interpretations of the speaker's intentions.

Understanding the characteristics of a good test, i.e. its validity, reliability and practicality, is the next competence required to be owned by test developers. In addition, test developers are also required to be able to prove the existence of the three characteristics of the developed test. Having understood these features, test developers need to follow the sequential steps in developing the listening comprehension test. The steps to be followed are: examining objectives and materials of listening materials in the syllabus, stipulating the intended ability to be measured, drawing table of test specification, selecting texts for the listening comprehension test, writing test items, refining the test, and

analyzing the test items.

For the betterment of achievement test of listening comprehension, the lecturers of listening subject are suggested to analyze the results of semester test of listening comprehension. By analyzing the scores of listening comprehension achievement test, the lecturers can gain useful information related to the reliability coefficient of the test, item difficulty index, and item discrimination index.

REFERENCES

Brown, J.D. (1996). *Testing in Language Programs*. London: Prentice Hall Regents.

Brown, H.D. (2001). *Teaching English by Principles: An Interactive Approach to Language Pedagogy*. New York: Longma.

Djiwandono, M.S. (1996). *Tes dalam Pengajaran Bahasa.* Bandung: Penerbit ITB Bandung.

Fulcher, G, and Davidson, F. (2003). *Language Testing and Assessment: An advanced resource Book*. New York: Routledge.

Grondlund, N.E. (1985). *Measurement and Evaluation in Teaching*. New York: Macmillan Publishing Company.

Heaton, J.B. (1991). *Writing English language Test*. London: Longman

Hughes, A. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Schoer, L.A. (1972). *Test Construction*. Boston: Allyn and Bacon.

Masden, H.S. (1983). *Techniques in Testing*. Oxford: Oxford University Press.

Morley, J. (1984). *Listening and Language Learning in ESL: Developing Self-Study Activities for Listening comprehension*. New York : Harcourt Brace Jovanovich, Inc.

Nunan, D. (1993). *Introducing Discourse Analysis*. London: Penguin Books Ltd.

Rost, I. (2001). *The Cambridge Guide to Teaching English to Speakers of Other Languages*. Carter, R. and Nunan, D. (Ed). Cambridge: Cambridge University Press.

Weir, C.J. (1990). *Communicative Language Testing*. New York: Prentice Hall International.

Weir, C.J. (1993). *Understanding and Developing Language Tests*. New York: Prentice Hall International English Language Testing.