Developing an English Proficiency Test for Kurume University: Listening Comprehension*

Chieko Kawauchi
Donna K. Stripling

1. Introduction

This study attempts to develop a test for assessing the language proficiency levels of students at Kurume University. The specific focus of this study is on listening comprehension.

Until very recently only the Grammar-Translation method was employed for teaching English in the secondary school system in Japan. This method concentrated on reading and grammar. However, today there is a great demand from the society as well as from the students for developing speaking and listening skills. Many high school students go to a JUKU or a so-called Conversation School in town beside their regular school and learn listening and speaking. In addition, every year hundreds of Japanese students go abroad for various programs ranging from a short-term summer study to a year of home stay.

In view of these demands the Education Ministry has developed a new curriculum designed to give high school students more practice in communication. However, the current university entrance examinations in Japan still reflect the old Grammar-Translation-based method of teaching. Since these tests have not been designed to assess the levels of listening comprehension and speaking ability of the students, these students, once they are admitted to a university, are placed into Oral English or English Conversation classes without regard to their listening comprehension or speaking abilities. Within a few years it is expected that students will begin to show wider disparity in their listening comprehesion and speaking levels, and sooner or later all language classes will be grouped by proficiency levels.

Considering these circumstances, we feel it is necessary to develop an English language proficiency test to assess the ability levels of the students at our university. And hopefully, based on the test the students will be placed in the appropriate class and given appropriate instruction.

In this study we will first give an overview of the previous research including reliability and validity on listening comprehension. Then, we will look at the types of listening tests and do a comparison of various tests that are on the market. Some of these tests will then be administered to our students. Finally, we will propose effective forms, types and content levels of a listening comprehension test that we feel is the most appropriate for the students at Kurume University.

2. Overview of Listening Tests

2.1. Components and strategy for listening

What is listening? What are the components in listening? What does a listener do? Rost (1991, pp. 3-4) indicates that necessary components are:

1) Discriminating between sounds
2) Recognizing words
3) Identifying grammatical groupings of words
4) Identifying 'pragmatic units'- expressions and sets of utterances which function as whole units to create meaning
5) Connecting linguistic cues to paralinguistic cues (intonation and stress) and non linguistic cues (gestures and relevant objects in the situation) in order to construct meaning
6) Using background knowledge (what we already know about the context and the form) and context (what has already been said) to predict and then to confirm meaning

7) Recalling important words and ideas

Rost summarizes these components into perception skills (1,2), analysis skills (3,4), and synthesis skills (5,6,7). Earlier forms of listening tests seem to have emphasized the individual component as seen in the Sony Perception Test and Seido Listening Test which focus on the discrimination of sounds. Recent listening tests like TOEFL, TOEIC, and JACET mainly focus on identifying grammatical grouping of words, but questions about communication-based listening such as identifying pragmatic units are also seen. Testing individual components such as a perception test is not preferred.

Various reasons are considered to be behind this tendency. First of all, there are variables, to some extent, in the way of pronouncing some words among native speakers of English. This is due to the regional and ethnic differences of English. For example, English spoken in England is slightly different from that of the United States. There are also dialect differences within the same English speaking country. Second, more and more non-native speakers of English are demonstrating their own ways of pronunciation within the range of comprehensibility. It seems that the range of optimal sounds for listening comprehension is becoming less strict as the population of speakers of English has increased. Consequently, it has become very difficult to define a word with one single accurate set of sounds. That sounds of individual words can be recovered from the context in which the target words are used has often been discussed. Failure to catch the so-called standard pronunciation of the word is not as serious as the failure to catch the meaning of a whole segment. People have become more interested in context-bound listening rather than context-free listening exercises such as sound discrimination.

We assume, here, that listening ability is a coordination of the component skills mentioned above. These components, which appear complicated, are something all listeners do unconsciously in the great amount of listening that they do in every day life. It might be considered that listening covers the essential portions of language ability that a person has. Listening is in this sense a cognitive or mental process.

Current teaching approaches emphasize communicative competence which focuses on communication-based and learner-centered teaching of a foreign language. Savignon (1983, 1984) discusses four types of competence —linguistic, sociolinguistic, discourse, and strategic competence— which learners of a foreign language must develop. In order to achieve successful listening, listeners must also have a listening strategy. Rost (1991, p. 5) introduces the following four types of strategies for sccessful listening.

a. Social strategy: How should I deal with this situation? What is my relationship to the speaker?

b. Linguistic strategy: What words should I pay attention to? What unknown words and expressions can I guess?

c. Goal strategy: How should I organise what I hear? How should I plan my response? What is my goal for listening?

d. Content strategy: Does this make sense in view of what I already know about the topic? What can I predict?

Successful listeners are able to combine these strategies to foster higher listening ability. However, poor listeners are able to employ one or two strategies only to attain partial understanding. For example, many learners of English as a Foreign Language (EFL) tend to focus on linguistic strategy, or activating language knowledge alone and are likely to ignore the content knowledge such as thinking about the situation. This is partly because EFL learners lack the exposure to spoken English and partly because they have limited time for listening practice.

2.2. Reliability in listening tests

Listening tests must be reliable and valid. Tests may be inaccurate or unreliable in the sense that repeated measures may have different results. These measures may also be invalid in the sense that other abilities are mixed. The characteristic of reliability is sometimes termed consistency, accuracy, dependability, or fairness of scores resulting from an administration of a particular test.

Various sources of measurement errors which are considered as threats are derived from testees, testers, test administration and question items themselves.

1) Threats from testees

Fluctuations in true scores are caused by maturation, further learning or forgetting; Temporary psychological or physiological changes are caused by fatigue, emotional disturbance, or practice effects.

2) Threats from testers

Intra-rater error variance is caused by the same rater's inconsistent judgment which results from a lack of standard criteria, fatigue, or personality factors. Inter-rater error variance is caused by two or more raters' difference of judgment in circumstances where there are a lack of detailed rating schedules and insufficient scoring sessions.

3) Threats from test administration

There are many potential sources of measurement errors such as differences in the clarity of instruction, the time of the test administration, the extent of tester interaction with testees, the prevention of cheating, the reporting of time remaining, etc. Moreover, environmental inconsistency also introduces measurement errors: Interruptions and distractions with one group of testees and not another. In the case of the listening test, typical inconsistencies occur when the listening part is presented at different rates or volumes for different groups of testees. Within the same room, there may be a difference in the listening quality of the different places where the testees sit. Detailed written guidelines for test administration must be supplied.

4) Threat from question items

a. Test length

The number of items in the test influences the reliability of the test. According to Henning (1987, p. 78), "little reliability is gained by adding more than 75 items. 75 items and 0.78 reliability may be thought of as the point of asymptote."

b. Item difficulty

When tests are excessively easy or difficult for a given group of testees, skewed scoring distribution will result. With skewed distributions, scores are unnaturally compacted at one end of the scoring continuum. As a result, it is difficult to distinguish among testees in their ability, showing low reliability.

The most important characteristic of an item to be accurately determined is its difficulty. Item difficulty is determined as the proportion of correct responses. In general, item difficulty is most appropriate when it approaches the mid-point (0.50) of the difficulty range. Rejection of items is considered when the item index is "less than .33 or exceeds .67" (Henning, 1987, p. 50).

c. Item Discriminability

One of the most powerful analyses used to discriminate between weak and strong testees in the ability tested is item discrimination. There are several ways to calculate item discrimination including recent computer progamming, but the more traditional way is as follows. First, separate the highest scoring group and the lowest scoring group from the entire sample on the basis of the total score on the test. Research often indicates that the optimal size of each group is 28% of the total sample. Then, each test item is calculated by using the following formula (Henning, 1987, pp. 51-52).

$$\text{Item Discriminability} = \frac{Hc}{Hc+Lc}$$

$Hc =$ the number of correct responses in the high group

$Lc =$ the number of correct responses in the low group

The higher the index, the better the discrimination. It is considered that an index of higher than 0.67 indicates reliable discrimination. Popham (1981, p. 298) also offers the guidline of item discrimination index, recommending over 0.4 ($ID=Hc-Lc$) which is equivalent to 0.67 ($ID=Hc/(Hc+Lc)$) as a very good item.

d. Correlation between tests

When administered under similar conditions, a test must produce consistent results as a whole. Consistency of results is necessary for reliable tests. There are three basic methods of estimating reliability between tests; 1) the correlation between test-retest scores, 2) the correlation of parallel tests, and 3) internal consistency method such as the split-half method, Kuder-Richardson formula 20 (KD20) or Kuder-Richardson 21 (KD21) (Hatch and Farhady, 1969, Brown, 1988).

2.3. Validity in listening tests

Major well known kinds of validity are content validity, concurrent validity, predictive validity and construct validity.

1) Content validity

The focus of content validity is on the adequacy of the sample, in other words, the content of whatever testers wish to measure. To obtain content validity, testers must carefully define the content so that test items will correspond to the content.

2) Concurrent validity

Concurrent validity is defined as the extent to which test performance is related to some other valued test of the same ability. This is an empirical, criterion-related validity. It usually involves recourse to mathematical formulae for the computation of validity coefficients. For example, scores of two different tests administered concurrently or within a few days are correlated using one of the formulae for the correlation coefficient, and the resultant correlation coefficient is reported as the concurrent validity coefficient.

There is a lot of research that reveals a high correlation between listening ability and various language tests such as cloze tests, writing tests, speaking tests, and overall language tests. For example, in Oller (1979) the correlation between the Listening Comprehension section of the TOEFL and the cloze test was 0.73. The essay scores correlated better with the Listening Comprehension than with oral interview and a variety of other tasks.

3) Predictive validity

If testers want to infer an individual's future probable performance, predicative validity is used. For instance, in the case of our listening test, we might correlate this test score with successive grade-point averages of final scores of English classes to obtain predictive validity of the listening test.

3. Types of Listening Comprehension Tests

The "ability to take a test involving spoken input is part language ability and part procedual ability. Language ability is probed in the interpretation of the input texts and test items stems... Response option types may be usefully categorized by the skill operation required of the test taker" (Rost, 1990, pp. 182-183). In this section the various types of input texts and test stem items are put into five categories— 1) phoneme discrimination tests, 2) tests of stress and intonation, 3) dictation, 4) tests using visual materials, and 5) tests using statements, short passages or dialogues.

3.1. Phoneme discrimination tests

In the beginning stages of language acquisition, learners will most likely have at least some difficulty in distinguishing between some sounds: for example "sea" and "she". If the learners' native language does not have an equivalent sound for a particular sound in the target language, that sound will be especially troublesome. While some learners at the university level in Japan may still have trouble with a few sounds, this level of testing is best exploited at the junior high and high school level. These types of items included matching the word that they hear with the correct word or correct picture out of a group choices.

3.2. Tests of stress and intonation

Learning the ability to recognize stress, intonation, rhythm and juncture are also very important at beginning as well as at the more advanced levels. Knowing these features is considered more important in communication than the ability to discriminate between phonemes. Types of items include listening for the main stress or inappropriate responses.

3.3. Dictation

Dictation is one of the more familiar ways to examine listening ability. Learners are expected to write single sentences or complete texts. In single sentence dictation, a set of single sentences is read usually three times. Testees respond by writing down as much of the sentences as they can.

Complete texts dictation involves the reading of a short, but whole passage. The passage is usually read three times. The first time it is read at normal speed. The second time it is divided into meaningful segments. Each segment is read at normal speed with a short pause between segments. The third time the segment is read at normal speed. Testees respond by writing down as much of the passage as they can.

3.4. Test using visual materials

Various kinds of visual materials such as pictures, photographs, and diagrams have come to be used in many proficiency tests. Some of the examples are selecting true/false pictures, or selecting correct picture or the most appropriate picture. For example, a simple diagram such as a series of squares is drawn. A single item is then placed in relation to each square. Statements are then read about the diagrams.

Another example is following instructions and locating things on a map. Some questions are used to assess the testees' ability to follow directions. Usually testees hear the passage about directions on the map and are asked to find out which place to go. Some questions ask students' ability to read a map and to locate things on the map.

3.5. Tests using statements and dialogues

These types of tests are designed to measure how well students understand short statements and dialogs. Most of the current proficiency tests use this kind of test. There are 6 main types of test; 1) sentence matching, 2) choosing the best paraphrase, 3) selecting correct response, 4) answering questions about a dialogue, 5) answering questions about a short passage, and 6) making inferences from the dialogue.

In sentence matching testees hear a sentence and then choose the sentence that matches exactly one that they heard. Choosing the best paraphrase and selecting the correct response ask testees to hear one sentence and then choose the sentence that most closely matches the original in meaning or the correct response. Answering questions about a dialogue or a short passage asks them to listen to short dialogue or a short passage and then answer a question about them. Testees have to select the correct response out of four written choices. In making inferences from the dialogue testees hear a short dialogue and then select a sentence that draws the best conclusion.

4. Comparison of Listening Tests

To develop a standardized listening test for the students of Kurume University, we looked at the listening section of major listening comprehension tests and proficiency tests on the market. Table 1 shows the components of the major tests available in Japan along with the number of items (#item), answer selections (#choices) and the amount of time (time) to be spent for the whole test.

Table 1
Components of Major Listening Tests

test	components	#items	#choices	time
BASIC	picture/photo description	10	4	35
	short responses	10	4	
	short dialogues	10	4	
	short passages	10	4	
STANDARD	true or false	20	4	35
	same of different	20	2	
	short passages	10	3	
SLEP	picture description	25	4	40
	one sentence statements	20	4	
	a dialog on directions	12	4	
	long dialogues	8	4	
TOEFL	one sentence statements	20	4	45
	short talks	15	4	
	long passage and dialogues	15	4	

TOEIC	picture description	20	4	45
	one sentence questions	30	3	
	short conversations	30	4	
	short passages	20	4	
Cambridge	picture selection	7	4	
	a short talk	6	4	
	fill in the blanks	10	4	
	Yes/No questions	5	2	
Eiken (pre-1st level)				
	a short dialogue	5	4	10
	a long passage	5	4	
Eiken (2nd level)				
	one sentence statements	5	4	5
	a short dialogue	5	4	

In our study the JACET standard (STANDARD hereafter), JACET basic (BASIC hereafter) and SLEP tests were selected.

The STANDARD test was chosen because it was the first test specially developed for Japanese college students. The test was produced by the Japan Association of College English Teachers (JACET) in 1975. Many college students take the test each year, and this will enable us to obtain not only the results of our students but also the overall information data of the nationwide results.

The BASIC test was developed as an easier version of the STANDARD. The level of the BASIC test is considered to be suitable for the students of Kurume University, because they are non-English majors and most of them are considered not used to listening.

Finally, we chose the SLEP (Secondary Level for English Proficiency) test. Many students hope to study at an American university, and more and more of them are taking TOEFL every year. The SLEP is an easier version of the TOEFL, and it provides equivalent scores of the TOEFL at the same time. The SLEP test can be given with a reasonable cost, and it can be repeated as often as desired. Also, we can obtain results on the spot. The TOEFL was not chosen because we felt that it was not appropriate for Kurume Students. The content and vocabulary level for the listening comprehension section used in the TOEFL seem too difficult for our students to reveal their real ability of English proficiency. Moreover, the administration of the TOEFL is costly and takes time to get the results.

The following is the brief explanation of the STANDARD, BASIC, and SLEP tests.

4.1. STANDARD test

The test was developed in 1975 for Japanese college students who have studied English as a foreign language. The main purpose of this test is to examine basic skills taught in the English classes in Japan. According to the manual, it attempts to exclude the cultural knowledge about the western world in order to be culturally fair. Stylistic and suprasegmental differences such as intonation and stress are also avoided. The test is not suitable a proficiency test for a specific purpose such as study abroad, but it is useful for examining basic ability of Japanese learners. The test has two forms, Form A and Form B, so it can be used as a progressive test as well as a proficiency test.

There are 50 items divided into three sections. The first and the second sections consist of 20 items each, and the third section 10 items. The first section asks testees to hear a two-sentence statement and choose whether they are true or false. In section two, testees hear two sentences and chose if these two sentences are the same or different in the meaning. The last section, a short passage, requires testees to hear a question about the passage to be heard in the beginning and then the passage itself, and the question is repeated before they choose the correct response out of three choices written. Before each section an example is given with Japanese instructions.

4.2. BASIC test

The BASIC test was developed as a junior version of STANDARD test in 1983 and started to be in use in 1988. The test is a 40 item, four-option multiple-choice test. There are four sections each of which consists of 10 items. These are shown in Table 1. Like the STANDARD the BASIC test is not likely to be outside the knowledge of the students and should present no problems relating to cultural bias. Before each section, an example is given with Japanese instruction.

The first section, selecting the correct statement, requires testees to understand and identify correctly a sentence describing a single picture stimulus. The test uses simple black line drawings, and the statements about the drawings are easily followed.

The second section, choosing the correct response, requires testees to understand a question and to choose the correct response. The questions and answers are easily followed.

The third section, answering questions about a dialogue, requires testees to understand a short dialogue followed by a question about the dialogue and then to select the best response to the question. There are ten questions in this section. The dialogues, questions, and answers are easily followed.

The fourth section, answering questions about a short passage, requires testees to understand a short passage followed by one or two questions and to select the correct response to the questions. There are 10 questions in this section. The passage, questions, and answers are easily followed.

4.3. SLEP test

The SLEP is a group administered test of English language proficiency. It is designed to assess the readiness of foreign students for English medium instruction at the secondary level. The test was developed as a version of the Test of English as a Foreign Language

(TOEFL) to use in the secondary schools. It was published by Educational Testing Service (ETS). It is used for selection or admission to private school or for placement into the public secondary schools (Stansfield, 1984).

Every attempt has been made to link the language used in the test to language that students would meet in an educational setting in the United States. It is a language proficiency measure based on the language that is "likely to be encountered by a student attending high school in the United States or an American type high school overseas" (Stansfield, 1984, p. 4). Therefore, it is in no way culture-free, and it may be asking too much of Japanese university students who are totally outside the American system to take a test based on this cultural norm.

The listening section of the test is a 75 item, four-option multiple-choice test. The time required to administer the listening section including listening the directions, doing sample items, and answering questions is 45 minutes. There are four types of test items types — selecting the correct statement, sentence mathching, locating things on a map and drawing inferences from the dialogue. No examples are given for any of the section. The instruction is English throughout the test.

The first section, selecting the correct statement, requires testees to understand and identify correctly a sentence describing a single picture stimulus. There are 25 questions in this section. The test uses black and white pictures, and except for a few culturally specific items the statements about the pictures are easily followed.

The second section, sentence matching, requires testees to match a sentence heard on tape with one of four sentences written in the test booklet. There are 20 items in this test section. These test item types are also easily understood by the students.

The third section, locating points on the map and then deciding which cars the speakers were in. There are 12 test items in this section. Following the directions involving compass points seem very difficult for our students.

The fourth section, drawing inferences from the dialog, is based on extended conversations. There are three topics in the related conversations and each topic includes two participants. There are 18 items in this section. The conversations deal with events that typically occur in an American school. This section contains quite a bit of vocabulary that is specific to an American school setting, and the situations themselves place much of the material outside the cultural norm for Japanese students.

5. Method of Testing

Subject

One hundred forty-two students from the departments of Medicine, Literature and Commerce participated in this study. All of the students were taking Oral English, or English concentrating on listening skills in the language laboratory. The medical students are all freshmen and have on the average 8.1 years of English learning. It is considered that they worked exceptionally hard studying for the competitive entrance examination to enter the medical school. The rest of the students are sophomores except for 10 freshmen who were taking Oral English. The average years of learning English for them is 6.9. None of them are English majors. One of the medical students lived in the United States until the second grade of school and has a fairly well developed command of English. However, the majority of the students had no experience of studying abroad.

Material

The STANDARD, BASIC, and SLEP tests were used. The components are explained in the previous chapter. Thirty-two medical students took the SLEP and BASIC tests, while 29 medical students took the

SLEP and STANDARD tests. Eighty-one students from the departments of Literature and Commerce took the BASIC test alone because of the limited time in their class schedule.

Procedure

All the tests were given in October, 1993 during the regular class period. They took the tests in the language laboratory without the prior notice.

6. Results and Discussion

The average percentages of correct responses of the STANDARD, BASIC, and SLEP tests are shown in Table 2.

Table 2
Average Percentages of Correct Responses of the Tests

	average(%)	mean	# items	highest	lowest	SD
STANDARD	60.4	30.2	50	48	17	6.6
BASIC	60.0	24.0	40	37	11	5.8
SLEP	66.8	50.1	75	70	39	6.9

The average percentages of correct responses suggest that there are no distinctive differences between these tests. However, it is hard to compare them fully because the scores of the BASIC, STANDARD and SLEP tests are expected to show originally in alloted points based on the respective weighting scales. The types of the questions in these tests are also different. We will analize the question items in the individual tests later in this paper.

In order to see reliability between the tests, the correlations between the parallel tests, SLEP/BASIC and SLEP/STANDARD were examined. There is a strong correlation between the SLEP and BASIC tests (r=.77, p<.01) and between the SLEP and STANDARD test (r=.71, p<.01). This means, for example, those who got higher scores on the SLEP tend to obtain higher scores on the BASIC, and also those who got lower scores on the SLEP tend to do so on the STANDARD. It also suggests that these tests are concurently valid, because they were administered within a week and resulted in high correlations.

Next, we examined the predictive validity: how the tests predict an individual's future probable performance such as the final grade point in the class. The correlations between these tests and the final grades of each student show blurring. There were moderate correlations between the BASIC test and the final grades ranging from r=.48 to r=.57 at p<.01 among classes. On the other hand, the correlation between the STANDARD test and the final grades was very low at less than r=.30. However, the upper 28% of the students in the STANDARD and BASIC tests indicated higher correlation between r=.50 and r=.68 among classes. No correlation was found in the lower 28% of the students between their final grades and all the tests.

Finally we examined the item difficulty and item discriminability of each test item in the STANDARD, BASIC, and SLEP tests. First, item difficulty was examined. Item difficulty is most appropriate when it approaches the mid-point (0.50) of the difficulty range. According to Henning (1987), the items were rejected when their index was less than 0.33 or exceeded 0.67 as inappropriate ones. Then, item discriminability (ID hereafter) was examined. ID is a powerful tool in the discrimination of weak and strong testees in the ability tested as discussed in the earlier section. First, we separated the highest scoring group and the lowest scoring group both of which consist 28% from the total subjects in each test. The ID index less than 0.67 was rejected as an unreliable one (Henning, 1987). Tables 3, 4, and 5 show the number of questions that fulfilled the requirements of item difficulty (#reliable), the number of total questions (#total), the rate of reliable items out of total (% rate), and the average ID of each part

Table 3
The Number of Reliable ID Items in STANDARD Test

	# reliable	# total	%rate	average ID
Part 1	0	20	0	0.58
Part 2	6	20	30	0.64
Part 3	6	10	60	0.72

Table 4
The Number of Reliable ID Items in BASIC Test

	# reliable	# total	%rate	average ID
Part 1	1	10	10	0.58
Part 2	5	10	50	0.69
Part 3	9	10	90	0.79
Part 4	8	10	80	0.73

Table 5
The Number of Reliable ID Items in SLEP Test

	# reliable	# total	%rate	average ID
Part 1	2	25	8	0.61
Part 2	0	20	0	0.53
Part 3	0	12	0	0.62
Part 4	1	18	6	0.61

in the STANDARD, BASIC, and SLEP tests, respectively.

It is clear that the BASIC test contains the highest number of items of high discriminability, while the SLEP test has the fewest, and the STANDARD is in the between. In this section we will focus on the STANDARD and BASIC tests. The SLEP test will be discussed later.

It can be fairly said that sections 1 of both the STANDARD (true/false) and BASIC (picture description) do not appear to discriminate between weak testees and strong testees. The only item which passed the requirement in the Basic test is as follows.

BASIC No. 9 (ID=0.84, picture description)

A. There are no clouds at all over Kyushu.

B. Shikoku is covered with clouds.

C. Honshu is entirely covered with thick clouds.

D. Only the eastern part of Hokkaido is cloudy.

This question requires the subjects to understand and identify correctly a sentence describing a single picture. The students must not only understand the sentence but also infer or read a weather map.

Sections 3 and 4, answering questions about a dialogue or a short passage, in the BASIC test are considered reliable question items in discriminating the listening ability of our students. Their average IDs are also very high. On the other hand, in the STANDARD test only part 3, answering questions about a short passage, can fulfill the requirements. Selecting true or false and same or different after listening to a short passage in the STANDARD seems very difficult for the students. Although there are only two choices, and there is a 50% possibility of chance performances, the number of correct responses was low.

The following are questions whose IDs exceed 0.80 in the STANDARD and BASIC tests. These questions are considered to have a strong discriminability for our subjects. There are 8 questions in the BASIC and two in the STANDARD.

BASIC item 15 (ID=0.83, Choosing a correct response)

I'm a stranger here. Could you tell me how I can get to Korakuen Stadium?

A. Yes, you could.

B. No, you couldn't.

C. No, I can't catch a taxi here.

D. Yes, take a No.5 bus.

BASIC item 18 (ID=0.88, Choosing a correct response)
Ah, here's a comfortable seat. Do you mind if I smoke?

A. No, not at all.
B. Sorry, I can't.
C. Yes, you do.
D. No, there isn't.

BASIC item 26 (ID=0.82, a short dialogue)
Woman: Excuse me, Driver. I'm looking for the City Hospital. Is this the bus for the City Hospital?
Driver: Yes, it is. You get off at Green Street.
Woman: Please tell me when we get to Green Street.
Driver: Okay.
Question: What is the woman doing?

A. She is getting off the bus.
B. She is walking on Green Street.
C. She is on her way to the hospital.
D. She is returning home from the hospital.

BASIC item 27 (ID=0.83, a short dialogue)

A. We are having very unusual weather these days.
B. Yes, indeed. It's very cold for this time of the year. Take care of yourself.
A. Thanks, I will. And you, too.

Question: Why do they have to take care of their health?

A. Because the weather is too dry.
B. Because the weather is too wet.
C. Because the weather is too high.
D. Because the weather is too low.

BASIC item 29 (ID=0.88, a short dialogue)

Jiro: Mrs. Johnson, could you pass me the salt, please?

Mrs. Johnson: Here you are, Jiro.

Jiro: Thank you. This is a really delicious dish, Mrs. Johnson.

Question: What did Mrs. Johnson do?

A. Jiro gave her the salt.

B. She asked Jiro to pass the salt.

C. She put some salt on her food.

D. She gave Jiro the salt.

BASIC item 30 (ID=0.92, a short dialogue)

John: Did you know I got a new job, Cathy?

Cathy: No, John, I didn't. How do you like it?

John: Oh, it's great.

Question: What are they talking about?

A. The new job John got.

B. The wonderful present John got.

C. The new job John is looking for.

D. The office John liked very much.

BASIC item 40 (ID=0.83, A short passage)

In many places in the world, people set off fireworks on special occasions. In the United States, they set them off on the Fourth of July, Independence Day. But in China, they set them off to celebrate the New Year, which usually begins in February.

Question: Why do Chinese people set off fireworks in February?

A. To celebrate Independence Day.

B. To celebrate someone's birthday.

C. To celebrate the Fourth of July.

D. To celebrate the New Year.

STANDARD item 8 (ID=0.81, Same or Different)

A. The heavy rain delayed the train.

B. The train arrived on time in spite of the heavy rain.

STANDARD item 16 (ID=0.82, Same or Different)

A. If Dick had had enough money, he could have gone to the concert.

B. Dick could go to the concert, because he had enough money.

Next, we will list the questions whose IDs are less than 0.50 which means the weaker testees performed better than the stronger testees. There is one item in the BASIC and three in the STANDARD.

BASIC item 5 (ID=0.49, picture selection)

A. Sacramento is west of San Francisco.

B. Sacramento is south of San Francisco.

C. Sacramento is northeast of San Francisco.

D. Sacramento is southeast of San Francisco.

STANDARD item 1 (ID=0.47, true or false)

We choose our clothes according to the occasion. When we attend ceremonies, we usually wear dark suits.

STANDARD item 11 (ID=0.44, true or false)

The climate of London is mild. The North Pole has a much colder climate.

STANDARD item 2 (ID=0.41, same or different)

A. John promised me to come here at five.

B. John promised me that he would come here at five.

True or false questions in the STANDARD test require not only understanding the meaning but also common knowledge about things in the real world. More specifically they require the common knowledge available to the people who were educated and have lived here in Japan. In this sense it is considered that these types of questions are not asking genuine listening comprehension. It seems very difficult for our students to read a map involving compass points such as north, south, west, and east.

Lastly, the SLEP test is discussed. This test was given only to the medical students. The first section, selecting the correct statement that fits a photograph, seems easy except for a few questions. Seventeen questions out of 25 had on average over 67% correct which exceeds the requirement of item difficulty. However, there are 5 questions whose IDs are over 0.67 but 3 of them were rejected because the item difficulty for two of them was less than 0.33 and one of them exceeded 0.67. The vocabulary includes words that are far beyond what a Japanese university student would be expected to know.

The second section, sentence matching, shows the lowest average IDs. Matching a sentence with one of four sentences written is considered less effective. The students might be able to detect the answers just listening to the sounds without understanding the meaning. While this type of question was very easy for the students, it seems less appropriate for testing purposes.

The third section, locating places on the map, is problematic for most of the students. As stated above, following the directions involving compass points seem very difficult. Questions about locating points on the map and deciding which cars the speakers were in are not the usual types of questions that Japanese learners see. The average percentage of the correct responses is 60.5. Only one item exceeds the ID requirement but failed the item difficulty index, showing 0.26.

The fourth section, selecting from a rather long conversation, is also considered difficult. The subjects had to follow three topics in the extended conversations, requiring longer attention to the dialog. Also quite a lot of the vocabulary is too advanced for our subjects. Although 7 out of 18 questions exceeded the ID index of 0.67, 4 out of them failed item difficulty of 0.33. The average percentage of the correct responses was 39%, revealing that this section was the most difficult.

7. Conclusion

Strong correlations between the SLEP and the STANDARD tests and between the SLEP and the Basic tests mean that these tests would be reliable as a proficiency test of our students. Those who obtained higher scores in the SLEP test also got higher scores in the STANDARD and BASIC tests. However, the low percentages of item discrimination for the STANDARD and the SLEP tests suggest they may be inappropriate as a placement test for the students. The latter two sections in the SLEP tests are also difficult, and some of the items are found to be problematic for Japanese learners of English. Listening to the directions involving compass points must be modified. The instructions of the SLEP are in English, and with EFL students this can also present a real problem. As a whole, we consider the BASIC test to be the most appropriate test for our students if the first section is improved. The questions using visual cues must be more carefully modified as to the consideration of content and vocabulary.

Students who got higher scores in any type of the STANDARD, BASIC, and SLEP tests are also considered to participate more actively in class and obtain good grade. Final grades in class are usually based on an accumulation of short tests in the limited areas, which are not necessarily connected with general proficiency. However, the strong relationship between the test scores and the final grades in class indicates that those who had higher scores on the listening comprehension tests were more successful learners in class. This fact suggests that if they are grouped by proficiency levels from the beginning of the classes and taught in more intensive ways according to their level of needs, they will probably benefit more from the class. No correlation between the test scores and the final grades was found with regard to lower scoring students. This fact is, in a way, encouraging for both teachers and students, because it suggests that those students who obtained lower scores are not necessarily poor students in class. It is considered that if those students are placed in an class appropriate to their needs, they will probably develop their listening skills to a higher degree.

Finally, based on the item analysis, we would like to suggest the most appropriate listening test for our students in terms of form, type, and content.

Concerning the forms, the number of items, and the length of the listening test are acute problelms. Henning (1987) says it is better to include as many as 75 items, but it would be difficult to administer a test this long in the limited time of class schedule. Moreover, it would also be difficult to maintain students' attention for much more than 30 minutes. Therefore, we recommend a test with 40 - 45 question items which can be completed within 30 minutes as seen in the BASIC test. The instructions for the test should be Japanese and not in English, which can invalidate the test from the standpoint of time as well as the flow of listening. In order to make the test less complicated and/or less boring, the types of the test need to be between 3 and 4. It is recommended the number of choices be four. If the number of choices were two, as seen in the true/false and same/different questions, the outcome would be threatened by chance performance. Four choices would be the optimal number as seen in many proficiency tests.

Concerning the types of questions, listening to a short dialog and passages seem to be the most appropriate question types. Questions about selecting a response that best fits a visual cue such as a picture or photograph have been used very frequently and are considered to be one of the best ways to measure listening ability. This is because the testees's performance is less dependent on other skills such as speaking, vocabulary and reading (Heaton, 1985). The low scores of this type in the BASIC and SLEP tests indicate that more careful attention must be paid on the content of the question. The TOEIC test uses this type of questions in a large part of its test. It would perhaps be of use to examine the picture questions used in the TOEIC to get better idea of how to best exploit this type of question for our students.

There are some types to be avoided. It would be better to avoid the question types which ask testees to choose true or false because they call on students to display common knowledge. There is also a 50% possibility of chance performance. Selecting a same or different response after listening to short statements is also better to be avoided. These types of questions could be modified into the question type of listening to a short passage. Matching sentence is also considered less appropriate because it is not necessarily listening comprehension.

Lastly, the content of the questions is the most essential part of the test. In order to write original test items, it might be a good idea to resort to the BASIC test as a sample. Especially question items whose IDs exceeded 0.80 would be considered to be the best model for making test questions. Here are some examples about asking for a response or paraphrase after listening to a short utterance:

Ex. 1. "I'm stranger here. Could you tell me how I can get to Kurume Station?"

Question: Choose the most appropriate response

A. Yes, you could.

B. No, you couldn't.

C. No, I can't catch a taxi here.

D. Yes, take a No.5 bus.

Ex. 2. "It took John a long time to find he couldn't fix my television."

Question: Choose the best paraphrase

A. After a long time John realized he was unable to fix my television.

B. John spent a long time fixing my television and at last he was successful.

C. John took a long time finding my television.

D. John searched for a long time, but he couldn't find my television.

Listening to a short dialogue about an appropriate topic must be also included. Long extended conversations with more than three speakers would perhaps be better avoided. Also, students would probably find dialogs between members of the opposite sex easier to understand. After listening to the dialogue a question usually asks why and what about the content. For examples:

Ex. 3. John: Would you like to go to a movie tonight?

Mary: I'm sorry, but I have to work tonight.

Question: Why can't Mary go to the movies with John?

A. She already has a date.

B. She has to take care of her sick mother.

C. She has to work tonight.

D. She doesn't like John.

A short passage with one or two questions about the passage should also be included. However, it's better to avoid questions involving compass points or even the street names which are very rarely used when following directions in Japanese situations. Again, the level of vocabulary and content in the BASIC test is appropriate one. One of the examples is as follows.

Ex. 4. Yoshi called his girlfriend, Junko. "I've got tickets for the concert," he told her. "I'll meet you at the station at 6:00. We'll have dinner and then go to the concert hall." Junko was very happy. But Yoshi didn't come until 6:30. They didn't come until 6:30. They didn't have enough time for dinner. They bought some sushi and hurried to the concert hall.

Question: What did they do just before they went to the concert hall?

A. They bought some sushi.

B. They waited in front of the concert hall.

C. They hurried to the station.

D. They ate some sushi.

In conclusion, a rough drawing of the tentative listening test format is shown in table 6.

Table 6
Listening Test Format for Kurume University

section	# items	question types	# choices	time
1	10	Picture descriptions	4	6
2	10	Short responses/paraphrases	4	6
3	10	Short dialogues	4	9
4	10	Short passages	4	9

References

Brown, James Dean. *Understanding Research in Second Language Learning*. Cambridge: Cambridge University Press, 1988.

Hetch, Evelyn and Hossein Farhardy. *Research Design and Statistics for Applied Linguistics*. Rowley: Newbury House Publishers, 1988.

Heaton, J. B. *Writing English Language Tests*. Ninth impression. London: Longman, 1985.

Henning, Grant. *A Guide to Language Testing.* Rowley: Newbury House Publishers, 1987.

Manual for JACET Basic Listening Comprehension Test. Tokyo: Kaitakusha, 1989.

Oller, John Jr. *Language Test at School.* London: Longman, 1979.

Popham, James W. *Modern Educational Measurement.* Englewood Cliff: Prentice-Hall, 1981.

Rost, Michael. *Listening in Action.* New York: Prentice Hall, 1991.

——. *Listening in Language Learning.* London: Longman, 1990.

Savignon, Sandra. *Communicative Competence: Theory and Classroom Practice.* Reading, Massachusetts: Addison-Wesley Publishing Company, 1983.

Savignon, Sandra and Margie S. Berns. *Initiatives in Communicative Language Teaching.* Reading, Massachusetts: Addison-Wesley Publishing Company.

SLEP Test Manual. Princeton: Educational Testing Service, 1991.

Stansfield, Charles. "Reliability and Validity of the Secondary Level English Proficiency Test." *System* 12 (1984) : 1-12.

Technical Manual for JACET Listening Comprehension Test Form B. Tokyo: Kaitakusha, 1980.

Note

* This study was granted Ishibashi Zaidan Scholarship in 1993.

Synopsis

This study attempts to develop a listening comprehesion test for assessing the language proficiency levels of students of Kurume University. An overview of the previous research including reliability and validity on listening comprehension is discussed and a comparison of various tests available on the market is made. Three kinds of English listening tests, the SLEP test, STANDARD test, and BASIC test were given to 142 students of Kurume University and the results were analized. There are strong correlations between the SLEP test and the BASIC test and between the SLEP test and the STANDARD test, suggesting they are reliable tests. In order to see the predictive validity, the correlations between the BASIC/STANDARD tests and the final grades of the individual student were examined. There are moderate correlations between the BASIC test and the final grades of the students, but almost no correlations were found between the STANDARD test and their final grades. Concerning the level of students, however, the upper 28% of the students in the STANDARD and BASIC tests indicated higher correlations with their final grades. For the lower 28% of the students in the BASIC and STANDARD tests, there was no correlation between the tests and their final grades. This suggests that those students who obtained lower scores in these tests are not necessarily the poorer students in the class. It is considered that if these students are placed in a class appropriate to their levels and needs, their listening skills will possibly be better developed. The average percentages of correct responses in the SLEP, STANDARD and BASIC tests did not show any significant differnces, but the item analysis in the individual test revealed noticeable results. A very few items in the STANDARD and SLEP tests met the criteria of item discriminability (ID) which requires an index of higher than .67. The BASIC TEST, on the other hand, showed the highest rate of items with reliable item discrimination. It can be said that the STANDARD and SLEP tests might be less appropriate to discriminate between higher and lower levels of our students. The BASIC test was, among of those we looked at, considered to be the most appropriate placement test available on the market. Finally, some suggestions as well as sample test questions on the most appropriate listening test for Kurume University students are provided in terms of form, type, and contents.