DOCUMENT RESUME

ED 395 449 FL 023 654

AUTHOR Kuroki, Kenichi
TITLE Achievement Testing: A Final Achievement Test Model for Japanese Junior High School Students.
PUB DATE Mar 96
NOTE 35p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Analysis; *English (Second Language); Foreign Countries; Junior High Schools; Junior High School Students; Language Proficiency; *Language Tests; Linguistic Theory; Standardized Tests; *Test Construction; Test Format; Test Items; Test Reliability; Test Use; Test Validity
IDENTIFIERS *Final Examinations; *Japan

ABSTRACT

This paper discusses the construction of language tests, particularly for English as a Second Language (ESL), that focus on language use in real situations. Linguistic theories that provide background for language test construction are reviewed, and application of those theories to ESL instruction in Japan is examined, with attention to the particular constraints of the English teaching environment there. The first chapter looks at the types and purposes of language tests, and the concepts of reliability, validity, practicality, and backwash effect. The second chapter offers an overview of English testing in Japan, including the pressure to excel on standardized tests and the types of test items currently used. The next chapter considers how more proficiency-oriented tests can be developed for the Japanese junior high school context, and how the theories presented in the first chapter may be relevant in Japan. Finally, test specifications and examples of communicative proficiency-oriented test items are presented. A brief bibliography is included. (MSE)

Achievement Testing

A final achievement test model for Japanese junior high school students

Kenichi Kuroki

Togo Junior High School

Miyazaki Prefecture

March, 1996

Introduction

A question that English teachers often ask themselves is, "Am I getting the results from my classroom tests that I expect after my teaching?". This question is particularly relevant when students' performance on achievement tests is not consistant with their class performance. This result is particularly disheartening when a student is actively involved in class activities but performs poorly on the test. All teachers and administrators would agree that students who demonstrate mastery of the curriculum performance objectives in class should score well on achievement tests. What, then, causes the discrepancy between apparent achievement and low achievement test scores?

First and possibly most simply, the activities and tasks in which students were involved in their classes are different from the tasks and activities on the achievement tests. An example would be a teacher emphasizing grammar in class and then testing skills such as writing or reading on the achievement test. Or, a teacher may emphasize writing and reading in class but then test grammatical knowledge on the test. Although it may seem obvious that any achievement test should require students to demonstrate what they have learned, in fact many teachers may be guilty of failing to test what they have actually taught.

The second and perhaps the more important reason for a discrepancy is that the format of the tests may not be consistent with the manner in which the particular skill was taught. For example, an inconsistency will arise if grammatical structures are taught primarily through translation but are tested through matching structure with the situations in which they are used. Another inconsistency would arise if a writing teacher encouraged his students to write as much as possible without worrying about accuracy but tested students on the basis of their accuracy in writing. As Barr-Harrison and Horwitz (1994) pointed out, "...whatever testing approach is used, it should not differ from that used in instruction." (p.190). That is, if the teacher uses the grammar-translation method in teaching, he should use a similar format in testing the student's knowledge of grammar. If the communicative method is used in the classroom, then the test should require that the student demonstrate communicative competence, too. This point is of critical significance for Japanese teachers of English since many teachers who focus on developing fluency in class construct achievement tests that focus on

accuracy. Such unsuitable tests ask students to demonstrate skills or knowledge which they have not been taught. The harmful effect of this is that students are likely to study English only for the purpose of taking tests and not for the communicative purpose of language.

In Japan, foreign language teaching methods have begun to change from the grammar-translation method to a more communication centered approach since the Ministry of Education revised the Course of Study in 1989. However, although the curriculum has changed, classroom tests have not been developed to evaluate students' actual performance using the target language. It is true that the challenges of creating new ways of testing communicative competence are great. For example, there are usually 30 to 40 students in one class. Some of the most important standardized tests, such as the university entrance exam, still remain grammar-focused. Additionally, Japanese students do not have an immidiate need to communicate in English. In spite of this, since Japanese students are expected to be able to communicate in English, as is reflected in the new curriculum, teachers will need to construct proficiency-oriented tests which emphasize language use. Furthermore, the criterion for evaluation on these tests should be the extent to which students can communicate in the target language.

The primary focus of this paper will be on the construction of language tests which focus on language use in real life situations. The background for the construction of such tests will include a discussion of linguistic theories and the application of those theories to English education in Japan with special attention to the constraints of the English teaching environment in Japan. In addition, with the inclusion of examples of test specifications and test items, the paper will be a useful reference for Japanese teachers of English who often struggle to construct tests that will evaluate students' communicative competence.

The theoritical context of language testing will be explained in Chapter I. Among the basic concepts of language testing included in this chapter are the types and purposes of language tests, validity, reliability, practicality, and backwash effect. Next, Chapter II gives an overview of English testing in Japan including the pressure for standardized tests. This chapter will also include a discussion of test items in current tests in Japan in terms of validity and washback effect. The following chapter will consider how more proficiency-oriented classroom tests can be developed for the Japanese junior high school setting. This discussion will consider how the theories presented

in Chapter I can be applied in Japan. Finally, Chapter IV will offer test specifications and test item examples of communicative proficiency-oriented tests discussed in Chapter III.

Chapter I : Theoretical contexts of language testing.

In preparation for considering language testing at the secondary school level in Japan, this chapter will review the basic concepts and principles of language testing in general. This discussion will create the theoritical context in which to assess the situation in Japan.

A. The uses of tests

According to Hughes (1989), there is no single test which will meet the needs of all testing situations. A test which is appropriate for one purpose may not be suitable in another situation. Therefore, when constructing any test, the user should first consider the purpose of the test and the use of scores.

There are two major classes of language tests, proficiency and achievement. Besides these two main uses which will be described below, language tests may also be used for purposes such as discovering candidates' strengths and weaknesses or placing students in the appropriate level in a particular language program.

1. Proficiency tests

As the name indicates, a proficiency test is designed to measure a candidate's language proficiency. A proficiency test is not based on a specified curriculum of study followed by examinees in the past, but rather tries to measure an examinee's general level of language mastery. The Test of English as a Foreign language (TOEFL), required for most non-native speakers of English applying to universities in the United States or Canada, is one of the most familiar examples of a proficiency test.

2. Achievement tests

While a proficiency test is not directly related to a particular course, an achievement test is designed to measure the degree of the candidates' achievement of the objectives in a particular course or curriculum. From the results of the achievement tests, examiners learn whether the candidate has

achieved the objectives of the particular curriculum or course. Tests constructed by teachers are usually achievement tests and are concerned with how much the student has learned from their course of study. Hughes (Ibid) defines achievement tests as folows, "In contrast to proficiency tests, achievement tests are directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives." (p.10).

B. Basic concepts of tests

What is it that test users expect from a test? In other words, what are the characteristics of a good language test? The criteria that Hughes (Ibid.) sets for a test or testing system are as follows:

> a [good] test or testing system ... will
> a)consistantly provide accurate measures of precisely the abilities in which we are interested;
> b)have a beneficial effect on teaching (in those cases where the tests are likely to influence teaching);
> c)be economical in terms of time and money. (p.6)

1. Validity

As mentioned above, tests should provide "accurate measures of precisely the abilities in which we are interested". A test can be said to be valid when it measures only what it is supposed to measure and nothing else. For example, if a test designed to measure students' listening ability requires candidates to write complete sentences in response to a question, the validity may be in question because such a test in fact measures not only candidates' listening ability but also their grammatical knowledge. Unfortunately, in Japan, too many teacher-made tests may not be valid because they do not measure students' English skills but only their knowledge of English grammar. This problem will be discussed in detail in Chapter II .

2. Reliability

Consistency of measurement is another important principle in testing. In order to be reliable, a test must provide consistent results when it is administered to the same student or group of students. According to Harris (1969), there are several factors which affect test reliability. First, the adequacy of the sampling of tasks is one of them. The more samples a test includes, the more reliable the test scores will be. The second factor is the test method. An objective test is more reliable than a subjective test because the scorer gives same score repeatedly for the same performance, or two or more scorers give the same score for the same performance. Multiple-choice type items are perfectly reliable and open-ended type items, such as compositions, tend to be less reliable. For example, essay tests may not be as reliable as objective tests since the results can easily be influenced by the order of scoring. An essay scored just after a poor essay tend to be more highly rated, and an essay scored immediately after a good essay tends to be marked more poorly than it may be in isolation.

3. Practicality

Every test should be economical, that is, should be easy and cheap to construct, administer and score. For example, teacher-made tests should not be time-consuming to administer. However, a test designed to measure speaking ability through individual interviews of one hour cannot be said to be practical because it is obviously impossible to devote so much time with individuals in large classes.

C. Types of tests

So far in this chapter, various uses of tests and some basic concepts of testing have been discussed. The discussion will now focus on types of tests.

1. Norm-referenced vs Criterion-referenced tests.

Tests can provide test users with two types of information about examinees. One type provides information about the examinee in relation to the other examinees. For example, student A is two points better than student

B. Or student C's score places him in the top ten per cent among all examinees. Tests which provide this kind of information are called Norm-referenced tests. In norm-referenced tests, a score is meaningful only in relation to other scores. The score does not provide information about specific abilities or performances and no one knows whether or not student A can perform specific tasks in the target language.

The other type of information that tests provide is whether the examinee can perform specific tasks in the language. The tests which produce this kind of information are said to be criterion-referenced tests. Criterion-referenced tests present tasks for examinees to demonstrate actual language performance. Therefore, the scores indicate whether the examinees can perform a certain language task successfully, and if so, how well. In short, an examinee's performance is measured according to the criteria or description of the level, not in comparison to other examinees' performance.

2. Formative vs Summative evaluation

Formative and summative evaluation are concerned with when the information is obtained and how it is used. Formative evaluation takes place throughout a curriculum so as to discover students' strengths and weaknesses during the learning process. The purpose of formative evaluation is to provide students with feedback on their learning. Yes/No questions or True/false questions which are asked of the students in the classroom are examples of formative evaluation.

In contrast to formative evaluation, summative evaluation is carried out at the end of a course or specified periods in the course for the purpose of grading or selecting students. The main concern is the extent to which students have achieved the goals in a course. Chapter tests or term end tests are considered to be examples of summative evaluation.

3. Discrete point vs Integrative tests

Hughes (Ibid.) describes the distinction between discrete point and integrative testing as follows:

> Dicrete point testing refers to the testing of one element at a time, item by item. This might involve, for example, a series of

> items each testing a particular grammatical structure. Integrative testing, by contrast, requires the candidate to combine many language elements in the completion of a task. (p.16)

That is, a discrete point test is one which tries to measure knowledge of language elements or skills separately, such as a grammar test or a listening test. On the other hands, an integrative test requires the candidates to demonstrate several skills or knowledge at the same time to complete one task. A cloze test is an integrative test because the candidates will have to use not only grammatical knowledge but also knowledge of vocabulary and reading skills in completing the task.

4. Objective vs Subjective tests

This refers to the method of scoring. Objective tests do not require judgement by the scorer in scoring. Multiple-choice items or true-false questions are examples of objective tests. Subjective tests, by contrast, are ones which require the scorer to make a judgement in scoring, such as in an essay test or an oral interview. The more subjective the scoring becomes, the less reliable the scores will be because different raters might give different scores on the same performance.

5. Closed-ended vs Open-ended items

While closed-ended items are ones which require the examinees to choose one correct answer from several alternatives, open-ended items require examinees to formulate their answers using extended language. To make the distinction of these item types clear, some characteristics of each type will be presented here.

Closed-ended items such as a multiple-choice or true-false items are usually easy to score because the scoring is completely objective: no scorer judgement is needed. Therefore, good multiple-choice items are perfectly reliable and economical. In contrast, most open-ended items such as writing a composition or responding in an oral interview tend to be less reliable because the rater must use judgement in determining the appropriateness of answers. However, Hughes (Ibid.) cautions us that too many multiple-choice items are not successful. "Common amongst these [problems] are: more than one

correct answer; no correct answer; there are clues in the options as to which is correct (for example the correct option may be different in length to others); ineffective distractors." (p.61)

D. Backwash

Backwash can be defined as the effect of testing on teaching and learning. The backwash effect may be harmful or beneficial. If a test which is intended to measure writing skills consists of only multiple-choice items, teaching tends to focus on practicing those items rather than practicing writing skills. If only an oral interview is used as the achievement test, the classes are likely to emphasize oral interactions. Valette (1994) points out that backwash occurs both at the program level and at the classroom level,

> The washback [backwash] from national or state tests is strongest on the teachers who organize their lesson plans so as to prepare their students to do well on the tests. ... The washback [backwash] is also commonly found at the classroom level. ... students put in their learning effort on those elements and those skills that will be covered on the test or that will count for their grade. At the student level, this type of washback [backwash] is often refered to as "studying for the test." (p.10)

Therefore, teachers need to construct tests which have beneficial backwash effect on teaching and learning. However, it is true that many Japanese teachers of English, through no fault of their own, are now suffering from a negative backwash effect from entrance examinations because, as mentioned above, those tests are still grammar focused and require students to translate. This problem will be discussed in the next chapter.

Chapter II : The overview of English language testing in Japan.

The primary purpose of this chapter is to evaluate current tests in Japan using as a model the standardized high school entrance examination of Miyazaki prefecture and classroom tests used in Togo Junior High School. However, before referring to the tests themselves, it is important to put the discussion in the context of the change of English language learning objectives after 1989 when a revised Course of Study was implemented by the Ministry of Education. What follows is a discussion of the revised objectives, and an analysis of test items in current tests, specifically a high school entrance exam and classroom test.

A. The change in philosophy of English education in Japan.

1. Before 1989.

According to Koike and Tanaka (1995), although the grammar-translation method had been the standard language teaching method used in the English classroom until the end of 1940's, an effort to make English more communicative was initiated by the Ministry of Education in the 1950's and 1960's. During this period, the Ministry proposed developing teaching materials based on students' interests, reducing class size to less than 40, and emphasizing speaking and listening. Nevertheless, since the audio-lingual approach was the prevalent teaching method during this period, most teachers concentrated on language manipulation activities such as pattern practice or dialogue memorization rather than on meaningful communicative activities. In addition, both teachers and students still tended to stick to the traditional methods which focused on translating reading materials and reading materials and reading texts for detailed comprehension because the university entrance exams stressed reading, translation and grammar.

2. The revision of the Course of Study in 1989.

The revision of the Course of Study (the National syllabus) made a huge impact not only on the methods used in the English classes but also on the educational environment that students and teachers faced, including the

textbooks used in class, the number of Assistant Language Teachers (ALTs) and the number of class hours that English was taught in junior high schools.

As Koike and Tanaka (Ibid.) reported about this revision, "... proposals requested reconsideration of the objectives for teaching English and encouraged a revision of teaching philosphy toward a more communication centered approach in secondary schools." (p.19). The revised Course of Study clearly states that English should be taught so that students can communicate using it. The following is the overall objective for teaching English anounced by The Ministry of Education in 1989.

> To develop students' basic abilities to understand a foreign language and express themselves in it, to foster a positive attitude toward communicating in it, and to deepen interest in language and culture, cultivating basic international understanding. (Underline added.).

B. Discussion of a sample of current tests

With the introduction of the revised, communicative curriculum, it follows that the examinations used, both classroom and standardized, should be constructed so as to measure student's abilities in relation to the curriculum objectives. This section will discuss two current tests in use in terms of reliability, validity and backwash effect. That is, test items will be examined in terms of whether they test what they profess to measure, that they are constructed so that they measure consistently, and that they affect classroom practice appropriately.

1. Standardized tests: The high school entrance examination as a model

While university entrance exams still remain grammar-focused, high school entrance examinations have changed drammatically since 1989. For example, more spoken language is being used in the tests. Direct translation is rarely required. Listening comprehension items are always included. However, there are some problems in terms of validity, reliability and backwash.

The examination given to the junior high school graduates in 1995 consisted of five sections intended to measure listening, reading and writing skills. Although the sections were not specified by skills, it appears from an analysis that they measure the following:

section name	(Actual area tested)	Item type	Number of items
1. Listening	(Micro skills)	Multiple-choice	10
2. Reading comprehension	(Grammar)	Multiple-choice	5
3. Grammar		Scrambled sentence	5
		Completion	3
		Matching	2
4. Integrative task	(grammar and reading)	Completion	5
writing	(grammar)	Composition	1
5. Reading comprehension	(detail reading)	Multiple-choice	3
		Matching	5
		Completion	1 /40

Administration time:
45 min.

Overall evaluation:

First, since this exam has only forty items, the reliability is likely quite low. In other words, the number of items is too low to produce consistent results with repeated administrations. Although it is obviously impossible for the examiner to cover all areas of ability since time is limited, s/he must select carefully the type of items to ensure an adequate representations of learning objectives. This is particularly important since important decisions, such as high school entrance in this case, are made on the basis of the examination results.

Second, the limited scope of language tasks also reduces validity. Since this exam heavily emphasizes testing students' grammatical knowledge, the student with greater grammatical knowledge may in fact be placed at a higher level than one with better communication skills. It can be said, therefore, that this test is not based on the objectives specified in the curriculum. For example, some teachers may think teaching listening in the class is a waste of time. Instead since the entrance exams cover grammar knowledge, they would tend to focus more on teaching grammar so as to prepare their students for the tests. It can be said that this examination, as a whole, probably has a strong negative backwash effect on teaching and learning. Bachman (1990) claims that "The consideration of test content is thus an important part of

both test development and test use. Demonstrating that a test is relevant to and covers a given area of content or ability is therefore a necessary part of validation." (p.244).

Item evaluation:

As mentioned above, problems in this test exist in terms of test reliability, validity, and backwash. Now an anylysis of selected items used in the entrance exam will be given.

(1) Listening comprehension

Q. Look at the pictures and give the letter of the picture being discribed.

(Examinees will hear)

This is in a house. We need it in a dark room. We need it when we study or read books.

One of the problems with this item type concerns test validity. In preparation for this discussion, however, it is necessary to review the listening objectives of the curriculum. The Course of Study defines the objectives in listening as follows; "To enable students to understand the speaker's intended message in simple spoken English passages, to develop proficiency in listening to English, and to foster a positive attitude toward English."

An analysis of this item shows that it focuses on micro-skills rather than on macro-skills. In other words, this item aims to test whether examinees can understand the meaning of each sentence including recognition of the vocabulary and grammatical structures used. To choose the correct picture, the examinees cannot miss any information included in each sentence. For

example, when they listen to A, the examinees have to understand the key words: "dark", "study" or "read books" and the grammar structure "..., when S+V+P.". Although testing micro-skills is important for diagnostic purposes, at the end of the course it would be more appropriate to test macro-skills such as listening for specific information or gist, following directions or following instructions. As Hughes (Ibid.) points out, "As far as proficiency tests are concerned, there has been a shift towards the view that since it is language skills that are usually of interest, then it is these which should be tested directly, not the abilities that seem to underlie them." (p.141).

Second, the language in this item is very artificial; it does not contain any elements which spoken language usually has, such as repeating information, pausing, redundancy, hesitation, etc. This lack of authenticity may lead students into trying to concentrate on every word in the text, which makes the task more difficult and artificial. Even the language in test items should be as authentic as possible because understanding the intended message is the goal students are supposed to achieve.

As a result of focusing too much on the listening task at the sentence level and the lack of authenticity, students may concentrate on every single word rather than on the general meaning of the speech when they listen to English. In addition, they may focus more on studying grammar rather than on practicing listening skills. That is, this item could have a negative backwash effect on students' learning.

(2) Testing writing

The writing objective in the Course of Study is "To enable students to express their ideas in simple written English passages, to develop proficiency in writing English, to foster a positive attitude toward writing.". An example of a writing question on the high school entrance exam is as follows:

> Q. Please introduce Miyazaki prefecture with three English sentences. (Miyazaki pref.)

Although the number of sentences examinees can use is limited to three, this task may seem on the surface to be a relatively less controlled writing task. Because, as Heaton (1990) suggests, "The only really satisfactory way to assess a student's ability to write is by means of a composition test."(p.105), this task might seem appropriate. However, there are problems

in terms of validity.

This writing task does not provide enough context for people to write sufficiently. If the communicative aspect of language is to be emphasized, the writing task should reflect real life writing activities. For this reason, the task should provide much more information about the context such as the person to whom the letter is supposed to be written and the purpose of the letter.

In addition, even if all examinees are familiar with their prefecture, there is a danger that this item may test other abilities, such as their creativity. It may also effect reliability because, in such a situation, a direct comparison between examinees' performance would be difficult. Considering backwash, students are likely to stick to translating from Japanese to English instead of producing a message in English. They would, therefore, focus on producing grammatically correct sentences.

To make this task more authentic, it is necessary for the examiner to think about the likely contexts in which people, especially young people, write. Possible text types would be letters, postcards, notes, or forms for junior high school graduates, and the reader could be a friend in an English speaking country or an ALT.

2. Classroom tests: The term-end test as a model

Next, an examination will be given of a term-end test administered to a 9th grade class at Togo Junior High School in 1995. This term-end test was an achievement test given at the end of the third year in junior high school and was constructed by the classroom teacher. Below is a general description of the test:

section	Item type	Number of items
1. Grammar	Completion	4
2. Grammar	Matching	4
3. Grammar	Matching	4
4. Pronuciation	True-False	4
5. Grammar	Completion	3
6. Writing	Scrambled sentence	3
7 Vocabulary	Translation	10

8 Grammar	Matching	6		
9 Reading	Cloze	8		
10.Reading	Translation	5	/51	Administration time: 50 min.

Overall evaluation:

As is clear from the chart above, with 21 out of 51 items devoted to grammar, this test focuses primarily on grammar. Indeed, the only abilities this test measures are the extent of grammar knowledge or the ability to memorize those rules they learned in class. Even in the writing and reading sections, all items can be answered if students know grammar rules. This test does not measure reading or writing skills which were practiced in the classroom. This test clearly shows us that the abilities we should be measuring are missing. In other words, this classroom test does not measure the curriculum objectives that the students are expected to achieve. More dangerously, the test results might show us what students have memorized from the course-book. Backwash is, therefore, clearly harmful.

Item evaluation:

(1)Reading

Write the best word to fill each blank.
Rowena: Hello. (1) is Rowena speaking. I (2) lost.
I'm now at Kitamachi bus stop. How can I get (3)
your house?
Kumi: Well, you'll see a tall yellow building at the corner.
(4) left there and walk about 500 meters. You'll
find our house (5) a fruit shop and a restaurant.
Rowena: I see. Thank you.
Kumi: You're (6). We're (7) for the Sukiyaki party.
Rowena: Great. I'm looking (8) to it.

Because this passage is taken from the course-book, this item does not truly measure students' reading ability. They may be able to fill in the blanks with correct words simply from memory. In addition, there are several possible answers for blanks 2 and 5. For these reasons, the reliability of

this item is questionable. If a passage that students had not read before was used, this cloze type item would succeed in measuring the student's language ability because, as Hughes (Ibid.) mentions, "... to respond to them [the items] successfully, more than grammatical ability is needed..." (p.65). However, we should remember that Hughes also mentioned that a cloze test should be used only when examiners need approximate information about an examinee's language ability because this item type measures overall ability rather than a specific skill. Regarding backwash, if students know that the passage used on the test is the same as one in the course-book, they may try to memorize passages from the book instead of really improving their reading skills.

(2)Writing

Complete one sentence using the following words.
1. don't, you, to, have, lunch, cook
2. know, when, I, do, to, homework, my
3. you, do, know, the, book, buy, to, where.

Although the intention is to test writing ability, this type of item tests grammar instead of writing. Grammatical knowledge is only one of several elements involved in writing. In other words, this is not a direct test but an indirect test, focusing on one of several underlying skills. This item is not valid to test the extent to which the student can express his/her ideas in English. In addition, because all questions deal with one structure, the student who can answer one of three questions can also answer the others. Moreover, the ability to answer these questions is not the same as the ability to use this structure. It only tests recognition of a grammar rule. The harmful backwash effect is that students are likely to focus on learning grammar.

Summary:

Since an entrance exam has a strong backwash effect on classroom testing, it is not surprising that both tests have led teachers to emphasize grammar, although the term-end test contains more questions on grammar than does the entrance exam. The characteristics of problems both tests have can be summarized as follows:

(1)small number of items

(2)emphasis on grammar

(3)sentence level rather than paragraph level discourse

(4)lack of context

(5)lack of authenticity

(6)harmful backwash

Factors which contribute to this test format and content include the (a) number of examinees, (b) lack of preparation time, and (c) lack of administration time. However, since we as teachers need to be concerned with test validity, reliability, and beneficial backwash from our tests it is critical for both administrators and classroom teachers to strive to improve the content and format of tests.

Chapter Ⅲ: Considerations in developing an end-of-term achievement test for Japanese junior high school students.

To improve test validity, reliability, and backwash effect of junior high school English tests in Japan, this chapter will deal with the issue of developing a comprehensive end-of-term achievement test used in Japanese junior high schools. The test discussed in this chapter is a model for a final acievement test given to third year junior high school students at the end of a three-year language course. The test would be constructed by classroom teachers to assess the extent to which students have achieved the curriculum objectives after three years of instruction.

A. A communicative model

The test items on the test should require students to complete communicative tasks. The test should emphasize more communicative aspects of language rather than manipulation of grammar. However, Terry (1986) describes the typical problems many achievement tests have as follows:

> With current interest in providing our students with "real language" practice in true-to-life situations, new textbooks, as well as our own changing instructional strategies, encourage students to use the language consistantly and immediately in activities ranging from purely mechanical to open-ended exercises as well as in spontaneous interaction. Our tests, up to now, however, have maintained a traditional, tightly controlled, essentially discrete-point item structure." (p.523).

In the Japanese context, for example, while examiners have made an effort to contextualize the test items, many students are simply asked to look at sentences and fill in blanks, and in scoring, only accuracy may be accepted as correct. The question at hand is whether or not a test can measure students' communicative competence in English in real life situations, and if we can, how and what should we measure on the achievement test?

It is now necessary to present the characteristics of communication which should be taken into account in the construction of communicative tests. The following are the general characteristics of communication as described by Fisher (1984).

a)First, it is self-evident that communication requires the actual use of language to send and receive messages.
b)At a minimal level of communicative performance, students must go beyond the mere manipulation of language forms void of semantic content and attend to the formulation of phrases and sentences which have the weight of referring to real world objects and actions.
c)... participants in a communication event engage in that event for specific purposes, usually to request information about some topic or react to a request for information about some topic or to react to a request for information."
d)... the course of a series of communicative exchanges in a developing communication event is only partially predictable."
e)... natural communication takes place in a specific and concrete context which enables the participants to identify and to react to pertinent sociolinguistic parameters, ... " (pp.13-14). Underline added

Newsham (1989) defines the characteristics of communicative tasks by summarizing the claims by Keith Morrow, " ... a communicative task requires interaction, unpredictable use of language, purposeful use of language, authentic language, and a context (who, whom, when, where, why)" (p.340). From these characteristics, the following elements are considered to be important criteria in constructing achievement tests for Japanese secondary students.

1. Interaction

The test tasks should require examinees to interact to some extent. For interaction, at least one other person must be involved in a task, although for written communication, one person need not be physically present. Additionally, a task should require the examinee to use language structures to send and receive messages in the exchange of information. For example, since most completion-type items in testing grammar do not indicate a person to whom the examinee is going to send the message using the structure, those items in fact do not test the use of language but only test examinees' grammar knowledge. Therefore, test items should require examinees to send or receive messages using their grammar knowledge even in testing grammar.

2. Unpredictable use of language

The test tasks should provide for a wide variety of answers. In natural communication, when people ask questions, they do not always expect only one answer. That is, communication is usually open-ended. Therefore, although some control is inevitable to ensure test reliability in every test, the test items should be as open-ended as possible.

Since testing unpredictable language use is difficult and presents the teacher with problems concerning practicality and reliability, it is important for teachers to think about balancing the number of items between open-ended and closed-ended, and teachers should be allowed to accept as correct unpredicted answers which are still acceptable. Harrison (1984) refers to the aspect of reliability in communicative tests and explains how to set limits to the language performance. "One of the traditional methods of providing this control is to make the student show understanding of meaning in one language by expressing it in another: that is, translation. Another is to offer him a closed set of responses, as in a multiple-choice test. In a communicative test, the control is provided by the context." (p.12).

3. Specific context

A test task should provide enough context for examinees to communicate in English. Who are the communicators? When and why are they involved in the speech act? Since contextual information provides students with a general idea of language functions, they will pay more attention to the communicative aspects of the language rather than the grammatical features. As pointed out in Chapter Two, presenting language without a concrete context leads students into focusing only on manipulating language elements or memorizing grammar rules. Therefore, providing a specific context is essential in communicative testing, although the specificity of the context should be determined according to the level of the students' proficiency.

4. Authenticity.

In communicative testing, tasks should be as authentic as possible. In other words, both texts and problems set in a task should reflect real life. Nunan (1989) defines authentic materials as "any material which has not been

specially produced for the purpose of language teaching".(p.54). Some examples of authentic materials would include letters, menus, hotel brochures, street maps, timetables, etc. Nunan also points out the importance of task authenticity, "... tasks could be analysed according to the extent to which they required learners to rehearse, in class, the sort of skilled behavior they might be expected to display in genuine communicative interaction outside the classroom." (p.59).

How, then should authenticity be determined on an achievement test? Authenticity for Japanese students should be determined according to their proficiency level and interest area. Although many researchers point out the importance of a needs analysis in determining authenticity, since Japanese secondary students do not have communicative needs, it is necessary for teachers to think of social interactions they are expected to encounter in real life and events they are interested in.

Seedhouse (1995) refers to the concept of the target speech community to analyze learners' needs in the General English classroom, "When attempting to cater for psychological needs it can be very useful to try to define the learners' target speech community, so that one can visualize what is being aimed at." (p.61). From my teaching experience, the target speech community can be defined as the community in which young people are socialize including the worlds of international travel and entertainment. That is, we assume that Japanese secondary school students will travel to English speaking countries. Within this context, they may experience activities such as exchanging letters, taking a plane, going to a restaurant, introducing Japanese school life, etc. The events which can occur in this target speech comunity may be used as the tasks on the achievement test.

B. Included skills

The final achievement test is designed to measure students' listening, reading and writing skills. Each item in the test emphasizes measuring macro-skills rather than micro-skills so as to prevent students from focusing too much on every sentence or word.

However, testing speaking is not included in the test since that presents teachers with problems in terms of reliability and practicality. Direct testing used to measure speaking is usually time-consuming, and is, therefore, not practical for large classes. An alternative method for measuring speaking would be through an on-going evaluation such as portfolio assessment or observation.

This test does not intend to measure students' grammatical knowledge in isolation, either. Students' grammatical knowledge should be assessed for diagnostic purposes in classes. It may be argued that testing grammar should be included in the achievement test so that students can get high scores on entrance exams. There may also be an assumption that knowing grammatical rules implies the ability to communicate in English. However, since this achievement test is designed for third year students after three years of instruction, the test objectives should be based on curriculum objectives rather than those language elements such as grammar or sentence patterns. In other words, the final achievement test should be designed to measure students' communicative skills, and this would have a direct beneficial backwash effect. The following is a general discription of the criteria in each skill.

1. Listening

Students will have to demonstrate their ability to understand various types of aural texts such as dialogues or narrative speech. Students are expected to show that they can use the following skills; 1)listening for gist, 2)listening for specific information, 3)identifying the emotional state of speaker from tone and intonation, and 4)understanding the various functions of speech.

2. Reading

Students will have to demonstrate their ability to read authentic textual and graphical or tabular reading materials such as magazines, letters or timetables. Students are expected to demonstrate the following abilities: 1)skimming for gist, 2)scanning for specific information, 3)identifying logical relationships between sentences in a paragraph, 4) identifying logical relationships between paragraphs in texts of three to five paragraphs.

3. Writing

Students will have to demonstrate their ability to write short notes or to create two to three paragraphs. Students are expected to demonstrate the following abilities: 1) write a short letter to a friend on a familiar topic, 2)write short answers to questions, 3)create a paragraph from individual sentences using cohesion to link sentences.

C. Test item types

Both closed-ended and open-ended items are used in the achievement test. Closed-ended items such as multiple-choice or true-false items are used to measure listening and reading skills. The marking is objective. Open-ended items such as composition or essay items are used to measure writing skills. The marking is subjective. In the Japanese context, since only one teacher scores a set of compositions, scoring should be based on various criteria such as accuracy, fluency or appropriateness rather than on a holistic approach. Harris (Ibid.) points out the following advantages of using essay or composition items in testing writing:

> 1.Composition tests require students to organize their own answers, expressed in their own words. Thus composition tests measure certain writing abilities. (e.g., ability to organize, relate, and weigh materials) more effectively than do objective tests.
> 2.Composition tests motivate students to improve their writing; conversely, if examinations do not require writing, many students will neglect the development of this skill.
> 3.Composition tests are much easier and quicker to prepare than objective tests, an importantadvantage to the busy classroom teacher. (p.69).

It should be noted that communicative qualities such as an interaction, a specific context or authenticity are provided in each item.

Chapter IV: Test specification

This end of term achievement test model consisits of four separately timed sections: Listening, reading (PART A), reading (PART B) and writing. Approximate testing time is 100 minutes for the 60 problems in the test. The test is designed to be administered in two consecutive 50 minute time period. Following is an overall design of the test.

Section	Skills	Item types	Number of items	
I (30 min.)	Listening (PART A) (PART B)	Multiple choice Completion	25-30	
II (20 min.)	Reading (PART A)	Multiple choice	15-20	
III (20 min.)	Reading (PART B)	Completion	10-15	
IV (30 min.)	Writing	Composition	2-3	Administration time 100 min.

A. Section I : Listening Time: 30 minutes

Section I is in two parts, PART A and PART B. PART A is a test of listening for gist in dialogues. Students will hear dialogues and answer the questions that follow.

Example.

Text:

WOMAN: What a nice puppy!
TEENAGER: Thanks.
WOMAN: What kind is it?
TEENAGER: Uh, it's a she, but she is not a special kind.
She's just a mixture.

(Adapted from Hynes and Baichman, 1989)

Question: What are the two people talking about?

a. the weather b. a dog c. clothes d. a friend

PART B is a test of listening for specific information. Students will hear information and be asked to fill in the blank to complete a form, or follow the directions. The Text type will include weather forcasts, phone messages, interviews, and radio program announcements.

Example.

Question: A weather forecaster is giving the forcast on the Friday evening news. Complete the form to write the predicted weather conditions for Saturday in Chicago.

Weather on Saturday

Weather conditions	Temperature
	High ( )F Low ( )F

Text:

Weather forecaster: Well, it's still raining here in Chicago, and it looks like the rain is going to continue through the weekend. It'll be rainy and chilly tomorrow. The outlook for Sunday-more rain and colder. The predicted high for tomorrow is forty-five degrees Fahrenheit, but the thermometer is expected to dip to the freezing point tomorrow night, with a temperature of thirty-two degrees. I'm afraid cold weather is on it's way! Chicago. Once again, continuing rain tonight through Sunday. Current temperature, thirty-eight degrees. And that winds up our weather report for this evening. This is Dave spellman. Have a good night, and if you are going out, don't forget your umbrella.

(Adapted from Schecter, 1981)

B. Section II: Reading (PART A) Time: 20 minutes

Section II is a test of reading comprehension. Students will read 2-3 passages and answer the questions that follow. Text types will include letters, diaries and magazine articles.

Example 1.

Dear John,

How are you? I miss seeing you at school. How would you like to come to visit us the weekend of July 15? You could come on Friday night and stay till Sunday afternoon.

We could ride horses at the stable down the road. And you could take a tennis lesson with me. Let me know if you can come. My dad says we could pick you up at your house.

Your friend

Rhonda

(Adapted from Otfinoski, 1993)

Q.1. This is a letter _____.

a. of invitation b. of thanks c. Accepting an invitation
d. asking for help

Q.2. Rhonda wants to _____ at the weekend of July 15.

a. play tennis with John
b. visit John's house
c. ride a horse with her dad
d. see John at school.

Example 2.

Levi Strauss made the first blue jeans in the 1850s for the California gold miners. Jeans were cheap and strong, so the miners liked them. But by the 1960's young people everywhere were in jeans. A new gold rush? No, they were popular again because they were still cheap and tough and not part of the world of high fashion. They were not part of the world of social class and competition.

Today the world is different. Jeans are different too. They can be expensive and very fashionable. Famous movie stars wear them, and so do

princesses. What would Levi Strauss think of that?

(Adapted from Harmer and Surguine, 1987)

Q.1. Who were the first jeans for?

a. movie stars b. miners c. young people d. princesses

Q.2. Why were jeans popular in the 1850?

a. Because they were fashionable.
b. Because they were different.
c. Because they were cheap.
d. Because they were famous.

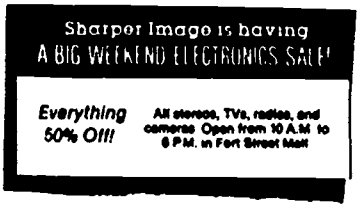
C. Section Ⅲ: Reading (PART B) Time: 20 minutes

PART B is a test of reading for specific information. Students will read passages and be asked to fill in the blank to complete a form, or follow the directions. Text type may include letters, advertisements, forms, timetables, plans, and street maps.

Example.

Read the following shopping advertisements to find the store which sells the following things. Check (✓) the store which sells them. Simpsons, Fisher, or Shaper Image? (Adapted from Richards, Hull and Proctor, 1990)

	Simpsons	Fisher	Sharper Image
bracelet	( )	( )	( )
CD player	( )	( )	( )
pencils	( )	( )	( )
sneakers	( )	( )	( )



D. Section Ⅳ: Writing Time: 20 minutes

Section Ⅳ is a test of writing skills including organizing and presenting information, and describing an object or event. Students will be asked to produce a piece of writing of the type specified such as notes, letters, or recipes.

Example 1

You were invited to Nancy's birthday party at six tommorrow evening. But you cannot go because you will be busy. Write a short note to tell her you cannot come and explain why. USE the information in the table to explain the reason.

YOUR AFTERNOON SCHEDULE	
5:30	ARRIVE AT HOME
6:00	JUKU
8:00	HOME
9:00	DINNER
10:00	HOMEWORK
12:00	GO TO BED

Example 2

Kazuo will leave Japan for the United State in two weeks. He has decided to write a letter about himself to his host family because the host family, Mr. and Mrs. Green, may want to know about him beforehand. Suppose that you are Kazuo. Write a letter using the following information.

Name	Kazuo Suzuki
Family	Father, Mother, Elder sister
School	Togo J.H.S.
Club	Baseball
Hobby	Listening to music
Favorite food	Hamburger
Girl friend	No

In the letter, you will have to 1) include all information above, 2) write more than FOUR sentences, and 3) write in the form of a personal letter.

Conclusion

The Development of appropriate achievement tests for Japanese secondary students is significant for Japanese teachers of English to consider. The primary concern in construction should be with the test validity and backwash effect. Lindquist (1951) claims that an achievement test, in its nature and purpose, should measure the specified educational objectives before actual construction in order to achieve good effects on educational practices. Therefore in constructing classroom achievement tests, the test constructors, classroom teachers in this case, should specify curriculum objectives, and the tests must be constructed to measure curriculum objectives.

In the Japanese context, however, although a communicative curriculum has been in effect in secondary schools since 1989, the achievement tests have failed to measure the curriculum objectives which focus on the communicative aspects of English language. In addition, as pointed out in Chapter Two, harmful backwash effects of entrance examinations focusing on English grammar or micro skills have extend to classroom testing as well as to students' learning. Current research in language testing suggests that test tasks on communicative tests should reflect real-life language use, specifically interactive, unpredictable, and purposeful uses, and both testing methods and texts used should be authentic. However, specific situations in Japanese secondary schools such as learners' proficiency level or lack of time for administration in fact has made constructing communicative achievement tests more difficult. Consequently, teachers may decide to test other easily tested outcomes such as knowledge of grammar rules.

This paper suggests a comprehensive model of achievement tests constructed by classroom teachers which emphasizes the communicative use of language. First, the test tasks require students to interact in the testing situation although the extent of interaction varies from one task to another. Second, the tasks also encompass unpredictable language use by avoiding extensive use of multiple-choice type questions and using an open-ended format for testing writing. Third, the test task provides a specific context for students to communicate in English requiring them to pay attention to language functions. Finally, the test tasks reflect real life performances and try to cover as many performances as possible that students may be expected to cope with. These include gaining information from both printed materials and non-text media or transmitting messages by writing.

As pointed out in Chapter Three, however, efforts should be made to measure the ability of spoken interaction, too. In order to cover a wide range of operations in real life, spoken interaction tasks will have to be developed in future tests. The alternative assessment to evaluate students' ability of oral interactions such as oral interviews or role plays is of considerable interest and futher research should be conducted in the Japanese context.

Bibliography

Bachman, L. F. (1990). Fundamental considerations in language testing. New York: Oxford University Press.

Barr-Harrison, P. & Horwitz, E. K. (1994). Affective considerations in developing language tests for secondary students. In Hancock, C. R. (Ed.), Teaching, testing, and assessment; Making the connection. Lincolnwood, Illinois: National Textbook Co. pp. 183-210.

Fisher, R. A. (1984). Testing written communicative competence in French. The Modern Language Journal, 68, 13-20.

Harmer, J. & Surguine, H. (1987). Coast to Coast: Student' book 1. New York: Longman.

Harris, D. P. (1969). Testing English as a second language. New York: McGraw-hill Book Company.

Harrison, A. (1984). Student-centered testing: Assessing communication in progress. (ERIC Document Reproduction No. Ed 273 113)

Heaton, J. B. (1990). Classroom testing. New York: Longman.

Hughes, A. (1989). Testing for language teachers. New York: Cambridge University Press.

Hynes, M. & Baichman, M. (1989). Breaking the ice: Basic communication strategies. New York: Longman.

Koike, I. & Tanaka, H. (1995). English in foreign language education policy in Japan: Toward the twenty-first century. World Englishes, 14(1), 13-25.

Lindquist, E. F. (1951). Prelimary considerations in objective test ccnstruction. In Lindquist, E. F. (Ed.), Educational measurement. Washington D.C.: American Council on Education. pp. 119-158.

Miyazaki Prefectural Board of Education (1995). Eigo gakuryoku kensa mondai.

Monbusho (1993). The course of study for lower secondary school foreign languages.

Newsham, G. S. (1989). Communicative testing and classroom teaching. The Canadian Modern Language Review, 45(2), 339-344.

Nunan, D. (1989). Designing tasks for the communicative classroom. New York: Cambridge University Press.

Otfinoski, S. (1993). Scholastic guides: Putting it in writing. New York: Scholastic Inc.

Richards, J. C., Hull J. & Proctor, S. (1990). interchange: English for international communication, student's book 1. New York: Cambridge University Press.

Schecter, s. (1984). Listening Tasks: For intermediate students of American English. New York: Cambridge University Press.

Seedhouse, P. (1995). Needs analysis and the General English classroom. ELT Journal, 49(1), 59-65.

Shizuoka Prefectural Board of Education. (1995). Eigo gakuryoku kensa mondai.

Terry, R. M. (1986). Testing the productive skills: A creative focus for hybrid achievement tests. Foreign Language Annals, 19(6), 521-529.

Vallete, R. M. (1994). Teaching, testing, and assessment; Conceptualizing the relationship. In Hancock, C. R. (Ed.), Teaching, testing, and assessment; Making the connection. Lincolnwood, Illinois: National Textbook Co. pp. 1-42.