Japanese University Entrance Examinations in English: Important Issues and Suggested Formats

David Peaty

Abstract

This paper considers key issues involved in the design and grading of English tests used in Japanese university entrance examinations (goals, target level, validity, reliability, feasibility, authenticity, etc.) and formats which might be used successfully for testing specific aspects of language proficiency.

Introduction

University entrance examinations dominate the Japanese education system as the eventual goal of 12 years' formal and supplementary education. They both reflect and influence what is taught in high schools and preparatory schools throughout Japan. They perform what is essentially a 'gatekeeping' function, enabling administrators to select a certain number of applicants on the basis of their overall test scores.

Because they have such a great effect on people's lives and because universities need to ensure that the most appropriate students are selected, it is essential that these tests measure as accurately as possible what ought to be measured and that they be graded fairly and consistently. TESOL specialists, however, often remark that many of the tests traditionally used by Japanese universities are neither valid nor reliable. Shillaw (1990), for example, claims that the typical translation subtests "lack construct validity .... and .... measure very little of the students' English proficiency".

My own research (unpublished) confirms the suspicion shared by many native speaker teachers of English that students who 'passed' reading tests which would confound many native speakers prove to be incapable of understanding far easier texts later in the classroom. These are clearly problems which need to be addressed.

Important considerations

Before we design a test, we must consider our *goals* carefully. Do we intend to select students on the basis of what they learned at school or of how well they will be able to follow and benefit from their freshman classes at university? These two approaches are very different and our choice should strongly influence our test design. The former approach requires an achievement test which must exclude anything that was not on the official curriculum and not used in the approved texts; the latter calls for a predictive test based on careful analysis of the skills needed in the first and subsequent years of university English courses. The former approach precludes the use of dictionaries, since we will test only what should have been learned; the latter, on the contrary, should encourage the use of dictionaries, since this is an important academic skill. The former approach would consist mainly of translation from English to Japanese and vice-versa, since this is how students were taught and tested at school; the latter would have to reflect the kinds of skills needed for university English classes and assignments, such as skimming, scanning, inferencing and note-taking. The extent to which the two approaches differ will depend, of course, on the extent to which university education differs from that of high school; if our classes are merely an extension of high school English, then the two approaches will lead to the same type of test. Hopefully, this will not be the case, however.

We also need to develop an idea of our *target level*. According to Buck (1988), one reason why reading tests are designed far beyond the actual reading competence of successful applicants is that "a test should look difficult, to give the impression that the college has a high standard". This has several bad effects. There is serious negative washback on the general education system, as a result of which high school students must struggle with highly academic and abstract texts when they have not even mastered the basics of English grammar and discourse. There is also a very bad effect on our test results, however; it drives scores down, resulting in a narrower spread

and a less reliable indication of relative proficiency levels. It is possible that many professors choose their test texts and questions on the basis of structural and lexical considerations. However, grammar and lexis represent only two of the many aspects of difficulty known to reading specialists (cognitive, discourse, stylistic, schematic, and so on).

One way to handle the difficulty problem is to place more emphasis on speed and less on difficulty. By increasing the amount of text to be processed and questions to be answered within a limited period, we can evaluate candidates' fluency and the degree of automaticity with which they can handle input without sacrificing accuracy, thus bringing our test closer to the proficiency levels we can realistically expect from six years of school English and the actual levels which will be demonstrated in our freshman classes. This would mean, of course, that tests would consume more paper or that less time would be provided, matters of concern to administrators.

A related consideration is whether we could and should use our test scores for subsequent placement of successful applicants in streamed classes. If so, then we need to be sure that our tests reflect the students' proficiency in specific areas (reading, listening, writing and speaking, if our classes focus on the macro-skills, and information processing if our classes focus on content). We also need to be certain that the test scores will indicate proficiency within the ranges in which we are interested. Regardless of the arguments of Oller (1976; 1979) and others in favour of unitary competence, experience (and TOEFL sub-test scores) tells us that many students are above-average in some macro-skills and below-average in others. Moreover, since students enter on the basis of total scores rather than how well they did in English, and many enter without even taking the standard test, there is bound to be a wide disparity between the competence levels of individual students. Among the advantages of using entrance examination scores for streaming decisions are that this would enable us to stream students roughly before classes begin and to choose textbooks at the appropriate level and that it would save time and labour otherwise spent preparing and conducting separate placement tests.

Key elements of testing

Having decided our goals and target level, we next have to confront the inevitable

conflicts between validity, reliability and feasibility.

Morrow (1979) summarizes the five main types of validity as follows: (1) face validity: "the test looks like a good one"; (2) content validity: "the test actually reflects the syllabus on which it is based"; (3) predictive validity: "the test actually predicts performance in some subsequent situation"; (4) concurrent validity: "the test gives similar results to existing tests which have already been validated"; and (5) construct validity: "the test reflects accurately the principles of a valid theory of language learning." If our test lacks face validity, it will surely be criticized by other members of the local teaching community and by the testees, even if it is otherwise an excellent test. This, unfortunately, tends to discourage innovation. We have already considered the need to choose between a test based on the official school curriculum, having content validity, and one intended to choose candidates who could cope with the demands of freshman classes in our institution, having predictive validity. Content validity, incidentally, now has another definition: "the degree to which the items in the test adequately reflect samples of the ability to be assessed" (Rost 1990: 177) If we try to establish concurrent (often called empirical or statistical) validity for our test before giving it, we run the risk of having questions leaked in advance, a disaster for any university; and in any case, cross validation is meaningless if we cannot demonstrate that the tests to which we are comparing ours are absolutely valid. We can, however, try to establish concurrent validity after the test results have been tabulated by, for example, comparing incoming students' test scores in the English section of our entrance exam with their scores on a TOEFL test taken immediately after entry. To expect a close correlation, however, would be somewhat optimistic. We can certainly attempt to identify the constructs underlying our test; but we would then have to justify them in terms of their pedagogical validity and relevance to our goals.

Our test must also be *reliable*. This means that all graders should assign the same grade to a single test paper when graded on a number of different occasions or, in other words, that graders share common standards and maintain those standards for all the papers they mark. This creates a potential conflict with content validity. If we wish to maintain an absolute standard, we need to design so-called objective tests, the answers to which are limited and pre-determined. However, such tests often fail to measure what they claim to measure.

The simplest objective test style is the True or False format. Multiple choice (usually from four alternatives) is also very convenient to score. Both types can be marked by computer and both are very common in Japanese university entrance examinations. One of the major problems associated with these formats is that they assess the ability to recognize a correct answer or to reject incorrect answers but not the ability to *produce* a correct answer. Moreover, since correct answers must be incontrovertibly correct and wrong answers must be decidedly wrong, answers tend to be either too obvious or too abstruse, often measuring the ability to decipher the question rather than comprehension of the text. The possibility of a lucky guess is also statistically significant: a fifty per cent chance in True or False questions and a twenty five per cent chance in four-item multiple choice. This possibility is often increased by the tendency of test designers to favour certain correct answer distributions and to avoid configurations such as 1-1-4-4-1-4, in which some numbers are not represented and some are repeated consecutively. The distribution of correction answers should be totally random. The lucky guess factor can be reduced in True/False question formats by adding a third option "not given or implied in the text", and in multiple choice formats by increasing the number of options; if there are no other plausible options, we can often include "all of the above" or "none of the above". A further problem is negative washback: students spend more time studying tests than studying the target language. The widely-recognized TOEFL Test suffers from all of these faults: the listening test reflects speed-reading ability and memory capacity; the whole test is based on the multiple-choice format, with all its defects; reading section questions are tricky enough to confuse native-speaker teachers who understand the texts perfectly; and test-taking strategies are nearly as important as language skills in producing high scores.

It is commonly assumed that subjective tests such as essays and oral interviews cannot be graded reliably and are thus unfair. In the pre-scientific era of testing, this was undoubtably true, and probably remains true with regard to the subjectively-graded translation sections of many Japanese university entrance examinations, which will be discussed later in this paper. Now, however, reliable grading of 'subjective' tests is quite common. For example, the T.W.E. (Test of Written English) accompanying the TOEFL is graded in bands according to estab-

lished criteria by graders who are so well-trained (after intensive two-day training programmes) that inter-rater reliability is almost 100%. This does not necessarily reflect the validity of the measures used, however; Brindley (1986: 21) warns us that "impressive reliability statistics for proficiency scales should be treated with caution".

The scoring of subjective tests inevitably conflicts with our third requirement, *feasibility*. In terms of manpower, time and space requirements, objective tests are far more convenient. It should be noted, incidentally, that there is not a clear dichotomy between subjectivity and objectivity in testing, but rather, a continuum. As explained above, essays can be graded according to pre-determined criteria with a high degree of reliability; on the other hand, cloze tests, which are normally thought to be objective, often turn up unexpected answers which some graders are prepared to accept but others are inclined to reject. In such cases, a decision has to be made and communicated to all graders immediately and all previously marked papers have to be re-checked, a very tiresome process.

Another factor to be considered is whether or not the texts to be used (printed or recorded) should be 'authentic'. This is partly related to the issue of whether we are testing what our applicants have learned or what will be expected of them; in the former case, our texts should be no more authentic than the ones used in high schools; in the latter case, they must be as authentic as the materials to be studied in freshman classes at university. If, however, we exploit authentic texts in an inauthentic way (e.g. by excluding essential contextual information or asking abstruse questions) then authenticity becomes meaningless. We must therefore take care to select texts which are self-sufficient or to include all relevant contextual information, such as titles and illustrations. We should also bear in mind that writers and speakers assume that their audience has certain background knowledge (of the topic, the native culture, the writer/speaker, life in general, and so on) which our testees may lack; unless testing such knowledge is one of our goals, we must ensure that our tests are not biased in favour of students who have such knowledge. A standard principle of test design is, of course, to avoid questions which can be answered without reading or hearing the text on which they are based. There are many other principles to be followed. Instructions should be clear, accompanied by examples and translated into the native language if possible - unless we intend to test the ability to understand abstruse

instructions. We must also avoid unnecessary cognitive burdens, answer sheets which, owing to poor design, lead students to write answers in the wrong place, and items which appeared in a similar context in our past tests, copies of which candidates may have used for their preparation. For further principles of this kind, the reader is recommended to look at Heaton (1988: 14).

We also have a social responsibility to maximize positive washback effects from our test. If we can include in our test elements which will enhance language learning at schools and preparatory institutions without making the test too difficult or unfair, we should include them. There is a tendency to avoid changing test formats even when they are proved to be bad, out of consideration for teachers and students of preparatory schools who have spent many hours working with past test papers. Such sentiments cannot justify bad test design.

Recommended test formats

In this section, we will consider various reading, listening, grammar, translation and communicative test formats. For further discussion and examples of various test formats, the following books are recommended: Grellet (1981) ; Heaton (1988) ; Hughes (1989) ; Madsen (1983) and Rost (1990).

Reading tests

Reading involves many skills at many levels. Our tests should reflect this by assessing both global and local comprehension and by requiring testees to guess, make inferences, read between the lines, understand irony and so on. Prabhu (1989) proposes a matrix in which questions may be at either local or global level and may require literal, inferential or evaluative reading. Following this matrix, we could ask questions such as:

a. (global/explicit) What is this passage about?

b. (global/inferential) Which country is the writer probably from?

c. (global/evaluative) Is the writer's style critical, sympathetic, polite or ironic?

d. (local/explicit) In what year did this event take place?

e. (local/inferential) In what season did this event probably take place? (implied by words such as 'freezing', 'overcoat', etc.)

f. (local/evaluative) What does the writer mean by 'ordered chaos'?

These questions could be incorporated into a multiple choice format.

e.g. This passage is about

1. the benefits of playing sport
2. the risks of injury from playing sport
3. types of sporting injuries
4. the treatment of sporting injuries

Prabhu also warns us to avoid questions which are harder to understand than the text itself. Such questions often result from the selection of a reading passage which lacks the ambiguities and nuances that test designers like to exploit. Texts must be chosen with great care; not only must they be the appropriate length, style and level of difficulty, they also must contain the potential for a good set of questions within the chosen format and be 'politically correct.'

Reading can also be tested by *cloze* formats. Of the various types, rational cloze is the most suitable. By targeting specific words for deletion, we can assess reading skills at various levels of discourse.

e.g.1 He works every day ..... Sundays.

e.g.2 Almost every gambler dreams of winning a fortune, but for most, it is a ..... that can never come true.

One problem with the cloze format is that there may be more than one possible answer, and more answers than we anticipated when providing graders with the answer sheet. Options can be narrowed down by supplying the first or last letter of the target word or by indicating the number of letters in the word, but this will lead to discrimination against testees who can understand the text well enough to sense the meaning of the deleted word but cannot supply the particular lexical item; in other words, this will become a test of vocabulary production as well as reading comprehension.

e.g. The panda r..... for its food supply on a rare species of bamboo which dies out every sixty years.

(This question penalizes the testee who knows 'depend' but not 'rely'.)

Multiple choice cloze is another possibility.

e.g. Expecting the manager to be an elderly gentleman, I was very surprised to see a young lady sitting at the .....A..... desk.

A 1. gentleman's 2 manager's 3 front 4 official

With this format too, we must take care to avoid testing lexical knowledge instead of reading proficiency; the options offered must be common words which all testees at our target proficiency level are expected to know.

A more complex variation is *cloze* summary. Testees read a text and then fill the blank spaces in a clozed summary of that text (or choose from a number of options if we have a multiple choice cloze format). This tests the ability to synthesize information from a reading passage; on the other hand, it may penalize testees who understand well enough but are poor at summarizing texts.

When dealing with small groups of testees with a common first language, *open-ended questions* asked and answered in the native language may be the most attractive format for a reading test in that it clearly tests reading and nothing else, allows us to target any skill at any level and provides testees with neither answers nor hints. However, it is not appropriate for large-scale testing as it creates both reliability and feasibility problems. These can be eliminated if we can design questions to elicit single-word responses; but such questions tend to be somewhat superficial.

Despite their defects, multiple-choice and multiple-choice cloze seem to enjoy wide acceptance here in Japan, along with the more dubious translation questions. If there is room for improvement, it probably lies in level, content and focus, rather than in changing test formats. Many university entrance tests of 'reading' are clearly tests of language processing instead, with texts which are extremely difficult and awkward questions focused mainly on explicit local meaning rather than global and inferential understanding. If translation tasks (which tend to focus on linear decoding and superficial meaning) are included elsewhere in the test, the reading sections should focus on deeper levels of discourse processing. If texts are too difficult linguistically, they should be simplified by lexical replacement or paraphrasing. If they are too difficult cognitively, they should be rejected. Questions should always be easier than the text itself.

There is a tendency to squeeze too many questions out of a single text. If an extract

does not have enough potential for the desired number of questions, it should be extended or rejected. There is also a tendency to restrict passages to a certain number of words. This may or may not be justified from the viewpoint of task difficulty (of which length is one of many factors) and of printing cost, but should not be allowed to prejudice the quality of the test.

Listening tests

Listening tests have been included in the entrance examinations of a number of Japanese universities in recent years. This should have a favourable washback effect on curricula and instruction at high schools provided tests are set at the appropriate level and well-constructed.

As with reading tests, *multiple* choice is one of the preferred formats. The listening section of the TOEFL test contains three parts, each designed for multiple choice. In Part One, testees listen to a single sentence and choose from four printed options the one most similar in meaning. This involves processing restatements and speed reading, in addition to listening comprehension. In Part Two, testees hear a brief conversation between two speakers, followed by a question about the dialogue by a third speaker, and choose the best from four printed answers. This tests the ability of the testee to remember the whole dialogue while listening to the question, in addition to testing speed reading of the four printed responses. In Part Three, there is a long lecture or conversation, followed by a series of oral questions about it. The testee is not permitted to take notes and has no idea what aspects of the talk the questions will focus on. This is therefore a memory test, even more than Part Two. Given the widely-recognised limitation of short-term memory to 7 items, this format is of very dubious value.

The listening section of the TOEFL is evidently not a pure test of listening. With certain changes, however, it could be useful for our purposes. To eliminate the speed reading component, we could either allow much more time between items and ensure that options are written in plain English or provide options written in Japanese. To reduce the significance of memory, we have to ensure that the sentences or conversations are short enough to be retained easily while the testee is processing the printed options, and that the question asked in Part B is provided in printed form (preferably

in Japanese). As for Part C, we must not only allow note-taking but also provide the questions in printed form (preferably in Japanese) in order to eliminate the need to memorise the entire speech. If we want to test comprehension of oral questions, we should provide a printed text (preferably in Japanese) to be read before listening to the questions. If we want to test comprehension of a lecture, we should provide an outline to be filled in or printed questions (in Japanese) to be answered by choosing from multiple options, also in Japanese, and allow note-taking. We should also bear in mind that authentic lectures include natural redundancy (repetition and rephrasing) and frequent pauses for note-taking; to deny our testees these authentic features would be unreasonable, in view of the other burdens we are placing on them (no visual cues, no advance notification of the topic, reduced sound quality, etc.).

One could argue in favour of a section in which the testee listens to a short utterance and then chooses which of four printed Japanese options best expresses what was said. This would eliminate the logical transformation component common in the TOEFL Part One in questions such as:

It's not unusual for Frank to forget his wife's birthday..

a. Frank often fails to remember his wife's birthday.

Buck (1988: 35) offers a number of criticisms of multiple choice tests, and argues that pre-testing is necessary and yet infeasible for security reasons, "which suggests that test-makers should avoid using multiple choice questions." Unfortunately, there are few alternatives for testing listening comprehension on many different levels and grading objectively a large number of test-takers.

Dictation is another popular format for testing listening comprehension (along with other aspects of language proficiency). Although it simply involves reproducing what was said rather than recognizing which of four options expresses the same meaning, it can be a very demanding task, depending on the speech rate and clarity, the length of pauses and pause units, the linguistic and cognitive difficulty of the text, and so on. It is generally believed that short-term memory constraints force the listener to process sentences in meaningful chunks rather than single words, and dictation effectively tests the ability to do this. However, the grading of dictation raises a number of problems related to how many marks to take off for different types of errors (those which alter the meaning as opposed to those which preserve it, for example).

Moreover, the grading of dictations is very tedious and time-consuming.

The *cloze* format may be used in listening as well as reading tests, using mechanical 'beeps' instead of white-ink to delete target words or phrases. A script cannot be provided, since otherwise this would become a reading test instead; therefore it is difficult for testees to match their answers on the answer sheet with the beeps on the tape. This problem can be avoided if we limit deletions to one per sentence and provide long pauses between sentences.

e.g. (Instructions in Japanese) Listen to the following sentences. In each sentence, one word or phrase has been replaced by a mechanical beep. Write down the missing words next to the numbers in your answer sheet. There are no penalties for spelling errors.

The candidates hear:

Number one. If you'd studied harder (beep) playing sports so much, you'd probably have passed the test.

The candidates should write:

1 instead of

As with reading cloze, there may be several acceptable answers. Options could be limited for easier marking by using a multiple-choice format

1. while 2. by 3. instead of 4. like.

One interesting variation is summary *cloze*. The testee listens to a short lecture or conversation (a lecture is more relevant to academic needs; eavesdropping is not of great value in university English classes) and then fills in blanks in a summary of the talk. This allows effective evaluation of the ability not only to understand holistically but also to synthesize input. The summary could be provided in Japanese. Multiple choice options could be provided for each blank if we wish to limit the number of correct alternatives for objective grading purposes; and these too could be provided in Japanese if we wish to eliminate the EFL reading component entirely. However, as Lewkowicz (1992) points out, the involvement of summarizing skills may detract from the reliability of our test as a measure of listening; its value lies mainly in its use for measuring students' ability to understand a lecture, take notes on it and synthesize it well enough to complete a cloze summary. Moreover, care must be taken to select or

prepare a well-structured talk which lends itself to effective summarizing.

Diagram and grid-completion tasks also offer possibilities for testing the ability to grasp the key points of a structured talk. They are somewhat limited in scope, however.

In designing listening tests, it is important to keep in mind the different cognitive operations which each type of format requires. For example, dictation requires listening and simultaneously processing, remembering and writing down; Part 3 of the TOEFL requires listening to the mini-lecture, processing and remembering almost every word in sequence, listening to each question, processing it, relating it to information provided in the mini-lecture, matching the number of the question with the set of four answers, reading and processing each answer, choosing the most appropriate one ....and all within strict time constraints. Generally speaking, we should aim at minimizing or simplifying the number of cognitive operations which are not directly concerned with listening. To this end, the use of Japanese in printed texts, questions or options will be very helpful, assuming that there is little difference between the testees as regards their native language reading proficiency. On the other hand, there is also the risk of negative washback if we appear to be encouraging simultaneous mental translation as an approach to effective listening comprehension.

Analysis of a recent listening test used in the entrance examination of a well-known Japanese university revealed the following problems.

1. Instructions were printed in English and were complex, possibly leading to errors unrelated to listening ability. The fact that many of the students have prepared for the test with the help of previous tests and therefore know the format justifies neither the retention of bad formats nor the use of complex English instructions; these merely bias the results against students with a high level of listening proficiency who have never gone to preparatory school.

2. Twenty of the forty-six questions involved choosing between two options, with a fifty per cent chance of a lucky guess.

3. Ten of the questions were based on the format used in TOEFL Part 2, and thus required various cognitive operations besides listening and reading.

4. The remaining sixteen questions were based on a modified version of the lecture format used in TOEFL Part 3. The lecture was played twice and note-taking was

encouraged. However, the lectures were of such an amorphous structure that useful notes could not be taken, and were based on oral readings of what was obviously written discourse, with no redundancy. As in the TOEFL, the questions were recorded, and worse still, so were the answers: the result being an extraordinary dependence on memory.

5. Some questions involved non-linguistic knowledge as well as listening or reading comprehension.

6. A number of questions were ambiguous, allowing several correct answers.

It is easy to offer belated criticism like this after a test has been conducted and published, but much more valuable to offer it at the design stage so that problems may be eliminated. It is therefore recommended that test designers work on separate tests and then take each other's tests with no prior encounter and under test conditions, since it may be impossible for even experienced test designers to recognize all of the flaws in their own tests. The designers of the test used in the previous year and those scheduled for the following year should also be involved in the checking process. Any item or format which proves difficult or impossible for a fellow designer should be rejected. This would, hopefully, eliminate problems of the kind mentioned above.

There is no simple, defect-free format for the objective assessment of listening; we should, therefore, use a variety of formats in order to spread the risk of interference. For further information on listening tests, please refer to Rost (1990: 175-221).

Grammar tests

The value of grammatical knowledge lies in its application in the creation or decoding of meaningful texts. If we test proficiency in any of the four macro-skills, therefore, we are also testing grammatical knowledge to a certain extent. The separate testing of grammatical knowledge would appear to be redundant. We may nevertheless want to test it for several reasons: firstly because our candidates spent six or more years learning it (i.e. for face validity), and secondly because it may provide some indication of testees' speaking and writing ability in the absence of more direct but unfeasible tests. Hughes (1989: 141-2) discusses in some detail the relevance of grammar tests and describes three formats: paraphrase, completion and modified cloze.

Paraphrase (also referred to as transformation) has been used in the Cambridge ex-

aminations for many years and is a good measure of testees' ability to express a given statement in a different way using a different structure.

e.g. Complete the second sentence so that it has the same meaning as the first.

A dog was chasing a cat.

A cat .....

Questions can be designed in such a way that only one answer is possible. In grading responses, we have to decide whether to give a single point for a completely correct answer or to deduct points from a predetermined maximum for each incorrect, omitted, wrongly included or misplaced word.

Completion involves filling in extended spaces. In ESL tests, this only works if the context clearly indicates what to put in the spaces and there are few alternatives.

e.g. It is a pity you could not come to the party. If you .....

....., I am sure you ..... enjoyed it.

With EFL tests, however, we can show exactly what should go in the spaces by providing a native language translation. We could supply either a translation of the whole sentence, to remove the reading comprehension element, or of the deleted segment. Possible alternatives could be limited by providing initial or final words.

e.g. If *it* ..... your help, we could never

.....the job on time.

Modified cloze can be designed to focus on specific grammatical features such as auxiliaries and prepositions.

e.g. I cannot speak Spanish but I wish I .....It ..... be very useful when I go to Mexico on business.

There are several other formats for testing grammar. The TOEFL test uses *multiple choice error identification.*

e.g. Choose the letter of the underlined word or group of words that is not correct.

A B C D

If I <u>had known</u> you <u>were coming</u>, I <u>had met</u> you <u>at the station.</u>

We need to make sure there is only one way of correcting the sentence.

Multiple choice identification of the *correct form* is another useful format.

e.g. If I had known you were coming, I .....you at the station. A will meet B will have met C would have met D had met.

Sentence combining tasks are also useful.

e.g. Combine these sentences without changing the meaning.

A child disappeared. The child's mother was very upset.

Sometimes there is more than one correct answer.

Translation tests

Heaton (1988: 18) cautions that tests of translation, "tend to be unreliable because of the complex nature of the various skills involved and the methods of scoring. In too many instances, the unrealistic expectations of examiners result in the setting of highly artificial sentences and literary texts for translation."

Despite these reservations, translation tests would appear to offer the only direct means of testing writing proficiency within our feasibility constraints. Heeding Heaton's warning, we must take care to choose relatively short, clear, context-neutral Japanese sentences that can be translated correctly in only a few different ways and to ensure a high standard of reliability in our grading.

During grading at a certain Japanese university recently, the following problems affecting reliability were observed.

1. Some Japanese graders (and even native speakers) failed to notice and penalize unnatural usage. Some native speakers applied the standard "Have I heard it used?", not realizing that they had only heard it used in Japan!

2. Some graders failed to match incorrect sentences with the closest correct model, thus deflating scores.

3. Other graders tried too hard to guess the intended meaning and gave the testee the benefit of the doubt, thus inflating scores.

4. Some graders used additive scoring (a point for each correct segment) while others used subtractive scoring (deducting a point for each error), which led to different standards.

5. Different graders applied different standards to errors of spelling, punctuation, article usage, word order, and so on.

6. Even native speakers failed to agree on how to treat usage which was archaic, usage which was incorrect but nevertheless commonly used by native speakers, and creative use of language that would be acceptable in poetry but odd in normal

contexts.

These problems can and must be dealt with in advance, by supplying graders with lists of correct options (including both American and British variants), examples of wrong answers with scores (to be used as models) and a clear statement of scoring policy with regard to different types of errors and whether additive or subtractive grading is to be used. The informal practice of having all graders at one table mark the same papers to compare the scores they gave and try to achieve inter-rater reliability must be formalized, and the score should be decided by a single appointed arbitrator whenever another scorer has any doubts. The number of acceptable variations should be limited by printing on the test paper below the segment to be translated a few English words at the beginning and end of each segment.

Communicative tests

The Japanese Ministry of Education has declared that from 1994 the high school English curriculum will be "communicative". This may mean anything from increasing the number of artificial dialogues in approved textbooks to a major directional change in favour of individual expression in writing and speech, depending on how seriously the proposal is taken by administrators, teachers and publishers. At any rate, universities may be able to contribute to this noble goal by making their entrance exams more communicative; and indeed they should do so if they want to select the candidates most capable of communicating with native-speaker teachers in their freshman year. (We need not concern ourselves with higher levels of communicative competence unless we are testing returnee and foreign students.)

Designing communicative tests for vast numbers of candidates within strict reliability and feasibility constraints is virtually impossible, however, for two of the four macro-skills. The Test of Written English (a supplement to the TOEFL test) requires two intensive days of grader training, as has already been mentioned; and this kind of training would appear to be inevitable for any reliable test of creative writing. The Test of Spoken English (another supplement to the TOEFL) is even more time-consuming and expensive since graders have to score recordings.

We are thus limited to the use of indirect but more practical measures of communication skills. Translation, discussed earlier, is one of these. The ability to ex-

press in English a series of propositions initially conceived in the native language is one aspect of communicative competence. Providing a set of propositions expressed in Japanese is one way of ensuring uniformity of task.

Grammatical competence is one of the four components of communicative competence defined by Canale and Swain (1980). Grammar tests can be designed to measure this. Socio-linguistic competence is another of the four components. This may be measured indirectly by testing the ability to produce an appropriate response.

e.g. Listen to the following questions and statements. After each, write an appropriate response.

Students hear "Thanks for your help" and are expected to write "You're welcome" or another appropriate response. Two points are given for each correct response; one point is deducted for each grammar mistake and for each spelling error that indicates a pronunciation mistake; two points are deducted for each answer that is socially inappropriate.

Grading could be simplified with the use of multiple choice.

e.g. Listen and choose the appropriate response.

1. Yes, of course. 2. Never mind. 3. You're welcome. 4. I'm glad.

This would then be a test of recognition rather than performance. Neither of these formats allow for phonological considerations; a candidate with incomprehensible pronunciation or inappropriate intonation could get all answers right and still not be able to communicate; however, they do measure knowledge and potential performance to some extent. Similar items could be designed for a test of written English.

e.g. (In Japanese) You are writing a letter to thank a friend for sending you a gift. Choose the most appropriate beginning.

1. Dear Eric, you kindly sent me a gift.
2. Dear Eric, your gift was very nice.
3. Dear Eric, I quite like the gift you sent me.
4. Dear Eric, I am writing to thank you for your gift.

Communication also requires the ability to produce appropriate lexical items. A cloze vocabulary test may therefore provide another useful measure.

e.g. I have a bad toothache. Can you recommend a good .....?

Naturally, low-frequency and archaic forms and expressions do not belong in any of

the formats discussed in this section.

It has been claimed (Oller; 1976, 1979) that certain reading and listening tests measure underlying language competence which is required for proficiency in each of the four macro-skills. If we accept a weak version of Oller's unitary competence hypothesis, we can use integrative tests such as cloze (for reading) and dictation (for listening) to indicate potential writing and speaking proficiency, although these may not be appropriate for decisions about placement of individual students in streamed classes focusing on writing or speaking. For a detailed discussion of unitary competence, please see Brindley (1986: 8-11), Brown (1987: 230) and Hughes (1989: 62-3). For further information on communicative testing in general, see Wesche (1983), Brindley (cited above), and Weir (1990)

Conclusion

The production and grading of English tests in university entrance examinations involve a huge expenditure of time, energy and money, yet the results are often very disappointing. More attention must be paid to goals, target levels, validity and reliability in grading, and more effective formats should be developed.

References

Brindley,G. The assessment of second language proficiency: issues and approaches. 1986. National Curriculum Resource Centre, Australia.

Brown, H.D. 1987. Principles of Language Learning and Teaching. Prentice Hall Regents.

Buck, G. 1988. Testing listening comprehension in Japanese university entrance examinations. JALT Journal 10.

Canale,M and Swain,M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1:1-47.

Grellet, F. 1981. Developing Reading skills. Cambridge University Press.

Heaton, J.B. 1988. Writing English Language Tests. Longman.

Hughes, A. 1989. Testing for Language Teachers. Cambridge University Press.

Lewcowicz, J.A. 1992. Testing listening comprehension using listening summary cloze. JALT Journal, Vol 14, No.1.

Madsen, H.S. 1983. Techniques in Testing. Oxford University Press.

Morrow, K.E. 1979. Communicative Language Testing: revolution or evolution. In Brumfitt, C.J. and Johnson, K.J. (eds.) The Communicative Approach to Language Teaching. Oxford University Press.

Oller, John W. 1976. A program for language testing research. Language Learning, Special Issue No.4.

Oller, John W. 1979. Language tests at school: a pragmatic approach. Longman Group Ltd. London.

Prabhu, N.1989. Workshop for Temple University of Japan.

Rost, M. 1990. Listening in Language Learning. Longman Group U.K. Ltd.

Shillaw, J. 1990. Entrance Examinations: Who Needs Em? JALT Journal, Vol.12, No.2

Spolsky,B. 1978. Introduction: linguists and language testers. In: B. Spolsky (ed.), Approaches to language testing. Arlington, VA: Center for Applied Lingnistics.

Weir,C.J.1990. Communicative Language Testing. Prentice Hall International.

Wesche, M.B. 1983. Communicative testing in a second language. In: Modern Language Journal 67: 41-55.