A Diagnostic Test for the Pronunciation of Native Japanese Learners with Regard to Certain English Speech Sounds

Timothy Ashton*

*Foreign Language Instructor, Fukuoka University

Abstract: At universities in Japan, under-graduate students are taught basic English conversation skills as a compulsory subject by native English-speaking teachers. Practice of pronunciation skills is often included as components of such courses. As these courses are typically subject to time restrictions, in turn limiting the time that may be allotted to pronunciation practice, teachers might benefit from more detailed prior knowledge as to which speech sounds their students find particularly problematic and would therefore be more worthy of inclusion. This paper will first identify which speech sounds in native spoken English would be appropriate to teaching groups of Japanese learners under such time restrictions. Secondly, a diagnostic test will be proposed, in order to investigate the strengths and weaknesses of the learners in their reception and production of these speech sounds. Thirdly, the construction of marking protocols and methods of
scaling to be used by examiners or markers, in order to evaluate the learners' abilities. Finally, there will be discussion regarding the reliability and validity of the test, factors that can affect this and strengths and weaknesses in its design.

1. Introduction: Statement of the Language Testing Problem

University English conversation lessons with a single teacher are typically held once a week for ninety minutes for one semester a year, with each semester on average lasting fourteen to fifteen weeks. Such classes typically consist of groups of twenty to forty mixed gender, native Japanese learners, aged between eighteen to twenty-two. Learners are of mixed levels of ability but come from reasonably homogeneous linguistic and educational backgrounds: generally, six years of exposure to English instruction in Japan at junior and senior high school levels but with minimal or no interaction with native English speakers. A few exceptional students may have benefited from occasional contact with native English-speaking instructors at high school or from participation in a home stay program abroad, e.g. in the USA or in Britain.

As regards course design for this subject, teachers are often left to their own devices, presenting and practising whatever linguistic items or skills they deem appropriate to their respective groups of learners, e.g. grammar, lexis and in many such cases, pronunciation.

The results of a diagnostic test, to be administered before or at the beginning of the course, could be used to determine which speech sounds need to be taught as high or low priority items during the course. Such
insight should aid the teacher, considering the time restraints and issues of priority in teaching items. In short, the aims of the test will be to assess the competence of the learners as a group and provide feedback for the benefit of the design of the ensuing course.

2. Considerations in Language Test Design: Bachman's Framework

The primary objective of language testing is to measure the language ability of a given individual. However, ability is in essence a mental quality which cannot be measured directly, as it is currently impossible to directly examine the brain and assess its capabilities. Language testing endeavours to do this in as appropriate and accurate a fashion as possible. Bachman (1990:40, 41) citing Thorndike and Haden (1977) provides a three-stage framework within which language tests may be constructed, linking the supposed ability, or 'construct', to the observance, and ultimately the evaluation, of its performance:

1. Identify and define the construct theoretically.
2. Define the construct operationally.
3. Establish procedures for quantifying observations.

2.1 Identifying And Defining the Construct Theoretically

The first stage of Bachman's framework states that the construct to be
measured must be distinguished from other similar constructs by defining it "clearly, precisely and unambiguously". Essentially, the test proposed within this paper will be used to measure the degree to which Japanese learners of English are able to articulate certain speech sounds in English and discriminate them when contrasted with (what native Japanese speakers usually perceive as) similar sounding English phonemes.² In other words, performance skills representing the mental construct termed 'pronunciation'.³

2.2 Defining the Speech Sounds to be Tested in Phonemic Terms

Rowe (1994) citing Nilsen and Nilsen (1971) lists a total of 32 different phonemes that native Japanese find problematic in an arrangement of 27 minimal/contrastive pairs. The list involves consonantal and vowel phonemes, including diphthongs and reduced 'schwa' [ə]. These are listed below:

1. /i:/ and /ɪ/ as in beat and bit
2. /eɪ/ and /e/ as in bait and bet
3. /æ/ and /æ/ as in aid and add
4. /æ/ and /ɒ/ as in cat and cot

² These items will be termed 'speech sounds' throughout discussion regarding articulation and as 'phonemes' when contrasted with others, illustrating their semantic properties.
³ The term 'pronunciation' will refer to the skills of aural recognition / discrimination of phonemes and oral articulation of speech sounds.
6. /æ/ and /ʌ/ as in bat and but
7. /ʌ/ and /ɒ/ as in cut and cot
8. /ʌ/ and /ʌ/ as in buck and book
9. /ɒ/ and /æʊ/ as in cot and coat
10. /ɒ/ and /ɔ:/ as in dot and caught
11. /ɒ/ and /æʊ/ as in dot and doubt
12. /ʌ/ and /ɔ:/ as in but and bought
13. /l/ and /r/ as in lack and rack
14. /w/ and /hw/ as in wet and whet
15. /v/ and /f/ as in vat and fat
16. /h/ and /f/ as in hat and fat
17. /b/ and /v/ as in ban and van
18. /v/ and /z/ as in veal and zeal
19. /s/ and /ʃ/ as in seat and sheet
20. /θ/ and /ʃ/ as in thank and shank
21. /t/ and /θ/ as in team and theme
22. /s/ and /θ/ as in sink and think
23. /ð/ and /z/ as in then and zen
24. /d/ and /ð/ as in dare and there
25. /d/ and /ʒ/ as in din and jin
26. /n/ and /ŋ/ as in fan and fang
27. /ŋ/ and /ŋ/ as in gag and gang

/ˌhw/ is included here to faithfully reproduce Nilsen and Nilsen's list. However, some teachers may not feel its inclusion necessary as it is generally a feature of British Received Pronunciation. Many speakers of standard Southern spoken British English articulate both 'wet' and 'whet' using /w/.
In their publication, Nilsen and Nilsen comprehensively chronicle the sounds in English that prove difficult for native speakers from various countries of the world, cataloguing them in a system of minimal/contrastive pairs, describing their phonetic and phonemic qualities and the communication problems that may arise through their confusion. As the concern here is with language testing and test design and not a contrastive analysis of English and Japanese Phonetics and Phonemics, it will be assumed that the authors’ claims are correct and accurate enough for the purposes of this paper and will not be disputed.

There are, however, points to be clarified in the use of the term 'minimal pair'. Minimal pairs are pairs of phonemes displaying only one allophonic variation from each other (Valette 1967:48) e.g. 19. / s / and / s / is an unvoiced alveolar fricative whereas / s / is an unvoiced palato-alveolar fricative. Some of the phonemic pairs listed above however, cannot be termed true minimal pairs as they display more than one allophonic variation, e.g. 16. / f / and / h / : / f / is an unvoiced labio-dental fricative whereas / h / is merely the voiceless counterpart of the following vowel. Similarly, 27. / g / and / η / : / g / is an unreleased, partially-voiced velar stop whereas / η / is a voiced velar nasal. In light of these differences, the term 'contrastive pair' will be used in preference to 'minimal pair' from here on.

2.3 Defining the Construct Operationally

Ability may only be measured indirectly (see 2) through observance of behaviour that represents it in an individual. Thus, the second stage of
Bachman's framework is brought into consideration: the construct must be isolated and made observable by means of a test, formatted to relate to the language behaviour that the construct is believed to represent (Bachman 1990:43, Davies 1990:52).

2.3.1 Diagnostic Testing

The ability 'pronunciation' is characterised through set habits in auditory reception and physical articulation by means of the organs associated with hearing and speech within an individual. It is this behaviour that must be elicited and tested in order to (indirectly) evaluate the ability (Lado 1961:42). A diagnostic test would be most suited to this purpose, for reasons summarised alongside the following basic criteria:

**Diagnostic tests are either theory or syllabus-based.** This test will be based on Nilsen and Nilsen's theories of speech sounds experienced as problematic by Japanese learners in English, as outlined in 2.2.

**Diagnostic tests are used to identify learners' specific strengths and weaknesses.** This test will investigate the realms of learners' reception and production of English speech sounds. Sometimes known as 'formative' or 'progress' tests, they may be used to check on students' progress in learning particular elements of a course. This particular test will be concerned with the overall progress to date that the learner has made in given aspects of English pronunciation but without reference to any specific method of teaching or study. In this respect the test may also be considered an 'achievement' test, making reference to the amount of control of the
language that the learner has gained. **The results of diagnostic tests may be used to determine appropriate remedial action to be taken to enhance the learners' skills**, e.g. in future courses and methods of teaching and learning with specific short-term objectives. The results obtained from this test will influence the content of the ensuing 14-15 week course, ultimately to enhance the students' overall performance in English conversation (Bachman 1960:60, Heaton 1975:165, Hughes 1989:13, Lado 1961:369, Underhill 1987:13).

### 2.3.2 Reasons for Testing both Recognition and Production

It is a general assumption that if a student can produce a speech sound, he or she will also be able to hear it. Similarly, if a student cannot produce a sound he/she will not be able to hear it. Lado (1961:41,78) stresses that this is certainly not the case however, and puts forward the following arguments (summarised):

(i) Recognition is usually more advanced than production, i.e. students learn to hear sound contrasts before they are able to produce them, probably because production requires more knowledge and mental processing than reception.

(ii) The differences between recognition and production skills are not the same for every student. Some learn to hear quite well but are unable to effectively articulate, others learn to articulate as well as they hear. Logically therefore, if only production were tested, it would not be discovered what the student were able to hear and vice versa. Recognition and Production must therefore be tested as separate sub-skills using
different techniques. In response to this, the diagnostic test should consist of two parts: Part I will be a phoneme discrimination test, to examine aural recognition and Part II will be designed to elicit speech sounds using visual stimuli. Descriptions of these parts of the test, along with a selection of sample test questions are outlined below:

Part I: Aural Test of Recognition and Discrimination of English Phonemes (Sample Questions):

Instructions: Tick ✓ the box □ next to the word that you hear on the tape. For example:
He has a problem with his □ vowels

□ bowels.
'Bowels' is the correct answer.

1. I don't believe it! I was □ hired this morning!

□ fired

2. Place your □ bats over there.

□ bets

3. I've got a dog called □ fang.

□ fag.

Part I of the test will essentially be a two-choice answer, listening comprehension test for the purpose of ascertaining whether the students are able to discriminate between phonemic contrasts, i.e. a sound contrast that changes meaning in English. Valette (1967:48) mentions that discrimination is one of the three areas within the skill of listening that require proficiency; the other two being retention and comprehension. Valette also
goes on to state that the ability to discriminate the sounds of a foreign language may initially be equated with the ability to distinguish the minimal pairs of the language.

The 27 contrastive pairs of phonemes listed in 2.2 should all be tested. These phonemes, when uttered in the context of sentences, contain troublesome contrasts for native Japanese learners. Each pair should be tested individually so there should be a minimum total of 27 test sentences involved.

**Method**: The students will hear a recording of the test sentences on a tape. These sentences will also be written on their answer sheets but with the test word containing the test phoneme paired up against a similarly pronounced word (the only difference in pronunciation being the test phoneme and its contrasting phoneme) but with a different meaning or spelling. These two words will be isolated from the rest of the sentence, with an empty box next to each. The students tick the box next to the word that they think they hear on the tape, i.e. a 'selected response' (Bachman 1990:129). The tape will only be played once with each sentence only uttered once. No indication will be made to the student (on the test sheet) as to which phoneme is being tested in each sentence as this would give away the answer.

**Part II : Oral Test of Production of English Speech Sounds (Sample Questions)**:

**Instructions**: Form sentences about your teacher using the pictures and words below. For example:
Sentence: "Tim is afraid of sharks"

1. late
2. beer
3. fat

Students usually learn to hear sound contrasts before they are able to produce them (see 2.3.2). However, Valette (1967:87) notes that learning to identify the phonemes of a new language does not of itself imply the ability to produce them. Therefore, a separate speech sound production test is required to determine the quality of a learner's articulation. This part of the test will utilise different techniques and be very different from Part I. Picture stimuli will be used to bring the student to utter a word containing one of the speech sounds to be tested in the context of a sentence. Bachman (1990:129) quoting Popham (1978) refers to this as a 'constructed' response, in which the response consists of the production of a language sample in response to input material.

Method: Individual students, in isolation from the rest of the class, will be confronted with a picture with a test word, containing the test speech sound, written underneath in order to provide a prompt to form a sentence
containing that word and consequently the speech sound. As there are 32 test speech sounds there should be at least 32 individual pictures with 32 test words. Each student will produce 32 sentences that will be recorded onto a cassette for grading later. This part of the test should take no longer than 20 minutes per student. In uniformity with Part I, no indication will be made to the student (on the test sheet) as to which speech sound is being tested in each case as this may hinder natural, unexaggerated articulation.

2.3.3 Criteria for the Designs of Parts I and II of the Test

Parts I (phoneme discrimination test) and Part II (speech sound production test) will vary from each other to some extent in terms of criteria: Part I will involve partially indirect testing, discrete point/item testing, two-choice answers and an objective marking system. Part II will also involve partially indirect and discrete point testing but will also employ criterion referencing and a subjective marking system with regard to the testees' answers.

**Indirect Testing (vs. Direct Testing)**: Bachman (1990:19) observes that an individual's physical attributes (e.g., weight, height) may be measured directly but mental abilities may only be observed indirectly. Consequently, all language testing contains an element of indirectness. Hughes (1989:14-16) further defines direct and indirect testing: testing is said to be 'direct' when it requires the candidate to perform precisely the skill which we wish to measure. 'Indirect' testing attempts to measure the abilities which underlie the skills in which we are interested.

Parts I and II of the text set out to examine the student's ability to
discriminate and produce a series of speech sounds in their behavioral manifestations, without the student's knowledge as to which ones he/she is being tested on, thus the testing is indirect. For it to be direct, e.g. in Part I, the student could be required to distinguish between two test speech sounds uttered in isolation, i.e. not within the context of a word. In Part II, the student could be asked to produce a certain speech sound in isolation. However, it should be noted that this test is not, in itself, entirely indirect. In Part I, the recognition of sounds is not totally artificial since they are in sentences but not fully contextualized either. There are deliberately no semantic clues to recognition in the test sentences so there is more emphasis on aural discrimination rather than recognition of elements of a phonemic system. Similarly, Part II also lacks full contextualization of the test speech sounds. Both parts of the test can therefore only be regarded as partially indirect.

As can also be understood from Hughes' definition, direct testing restricts itself to the mere observance of the skill, whereas indirect testing allows speculation on mental constructs and so is more suited to this purpose. Unfortunately however, due to lack of current knowledge of these constructs, it can only accommodate speculation.

**Discrete Point Testing**: The speech sounds listed in 2.2 will be examined individually, item by item, in both parts I and II just as they have been identified. This is an example of Discrete Point/Item testing (Harrison 1983:49, Hughes 1989:16, Bachman 1990:128).

**Criterion Referencing (Part II)**: Criterion referenced testing aims to interpret an individual's performance with reference to a specified criteria or a specific content domain (Bachman 1990:8) the 'domain' in this case
being the quality of articulation of certain speech sounds. Unlike norm-referenced testing, it does not discriminate between learners with the view to placing them in a rank order. Davies (1990:18) citing Valette and Disick (1972) proposes a model for criterion referencing, stating that the results of such a test be used to (i) give a relevant and non-subjectively defined learning programme and (ii) provide an individualised instruction programme.

It is hoped that the results obtained from this test may be objectively used by the course designer to (i) determine which of the speech sounds tested need to be treated as priority teaching and practice items during the time-restricted 14-week course and (ii) streamline the course to suit the requirements of the individual class.

2.4 Establishing Procedures for Quantifying Observations

The final stage of Bachman's framework requires the definition of methods of scaling and evaluating observations of the behavioral skills that the mental construct is presumed to underlie, i.e. the construction of marking protocols, keys and categories to be used by the examiners or markers. Parts I and II will both require individual objective and subjective scoring systems⁵ in order for the results to be interpreted reliably and validly (see also 3.3.2).

⁵ It should be noted that 'objective' and 'subjective' are terms referring to the scoring system only and not the format of the test itself.
2.4.1 Scoring of Part I (Aural Recognition)

Part I of the test will be scored objectively. Many of the authors referred to in this paper define 'objective scoring' as involving a single correct answer, unambiguously identified as either 'right' or 'wrong'. Consequently, no evaluative judgement is required of the marker and the test may be marked mechanically. The correct answers are then simply added up into a score. In the case of this part of the test, the listener will either recognise the phoneme or not and indicate thus on the test paper. There will be no need for further speculation.

2.4.2 Scoring of Part II (Oral Production)

As Part II of the test is criterion referenced (see 2.3.3) it will be scored subjectively. The common definition of 'subjective scoring' is that which requires the judgement of the marker, the opposite of objective scoring. A nominal scale will be employed, involving no 'right' or 'wrong' answers but a choice of two defined grades, 'satisfactory' or 'poor', falling at points (as Davies 1990:19 citing Glaser 1963:519 puts it) "on a continuum of knowledge acquisition ranging from no proficiency at all to perfect proficiency" (see fig.1). 'Satisfactory' and 'Poor' may be defined as follows (adopted from Lado 1961:80):

'Satisfactory': Articulation of the speech sound approximates native speaker delivery to the extent that it does not mislead the listener as to the meaning of the word it appears in.

'Poor': The speech sound is inaccurately articulated, misleading the listener.
as to the meaning of the word or rendering it unintelligible.

Fig. 1: The grades 'satisfactory' and 'poor' shown on a continuum of knowledge acquisition ranging from no proficiency at all to perfect proficiency.

As Harrison (1983:22) mentions, "the system of subjective marking depends on the aims and content of the test". The definitions of the grades contain no reference to 'native speaker-like' quality of production, which would be inappropriate as the students described (see 1) would not be capable of such precision. The grade 'satisfactory' establishes a suitable dichotomy and a 'cut off' (Davies 1990:18) so that non-native like production may still be regarded as successful or adequate. Obviously, a more detailed phonetic analysis would be more desirable in order to more finely discriminate the quality of articulation, but it would be more difficult for the marker to apply accurately by ear alone without phonetic training or perhaps the use of a spectrograph (Bachman 1990:44, Hughes 1989:19, Harrison 1983:110-111, Lado 1961:28-29, Underhill 1987:8, 94-96, Heaton 1975:11).

2.4.3 Combining the Results of Parts I and II

The two-grade scoring systems of Parts I and II ('right/satisfactory' and
'wrong/poor') correspond, making easy comparison and combination of results. The scores from both parts can be added together to form single scores. Those speech sounds scoring high numbers of 'wrong/poor' marks would be taught explicitly as priority items during the course whereas those scoring high 'right/satisfactory' marks would receive less attention.

3 Reliability and Validity

In order to provide an answer to the question 'Does the test work?' factors concerning reliability and validity of the test need to be taken into consideration. Regarded as both essential to test design, reliability has been described as "concerned with ensuring that a test is an accurate measure" whereas validity "attempts to provide a theoretical framework which gives reassurance to the test" (Davies 1990:6). Closely related, they are recognised as "complimentary aspects of a common concern in measurement" (Bachman 1990:160) and referred to again by Davies (1990:8) as "two poles, forcing a super-ordinate polar tension" upon the structure of the test, constantly requiring compromise. Reliability has also been described as a specific type of, or essential ingredient to, general validity: a test cannot be generally valid unless it is reliable (Hughes 1989:7, Heaton 1975: 155, Underhill 1987:105).

3.1 Aspects of Reliability

Reliability is a general consideration rather than specific. A 'reliable' test can be judged by the following criteria: a test generating scores which do
not fluctuate significantly, or for no apparent reason, if used on the same candidates repeatedly provided that there has been no intervening instruction (Bachman 1990:24, Hughes 1989:7, Underhill 1987:9, Lado 1961:31, 330). Also, a test that elicits language performance in a standard way, under uniform conditions (Bachman 1990:43). There are several intrinsic and extrinsic aspects concerning reliability which may yield elements of unreliability. These can be failures of the test itself, the circumstances in which the test is taken, the way it is scored and the uniformity of the assessment it makes (Harrison 1983:10, Heaton 1975:155-156, Valette 1967:31-32, Hughes 1989:3).

3.2 Aspects of Validity

Validity is a specific consideration as opposed to a general one. It depends on the linguistic content of the test and on the situation or technique used to test this content, resulting in the need to pose two closely related questions: does the test measure what it is intended to measure and is the test relevant to what it claims to measure? (Lado 1961:30, 321, Underhill 1987:9, Harrison 1983:11, Heaton 1975:10.) Certain independently definable aspects of validity are relevant to this particular test:

Content Validity: Reliant on the professional judgement of the teacher/tester, content validity raises the question of whether the test items and tasks cover the range of all the items that they were designed to test (Underhill 1987:106, Davies 1990:23, Lado 1961:322). Care must be taken to include all the target speech sounds listed in 2.2 at least once in both Parts I and II of the test.
Construct Validity: Construct Validity assumes the existence of certain learning theories and constructs underlying behaviour and skills and so questions whether the test represents the theoretical construct (and in turn the behaviour) that it is supposed to measure. Part I of the test examines aural recognition of phonemes by means of a phoneme discrimination test. Parts II tests oral production of speech sounds by means of visual stimuli. Phoneme discrimination and speech sound articulation are behavioral/performance skills representing the mental construct ‘pronunciation’ (see 2.1, Heaton 1975:154, Davies 1990:23, Underhill 1987:106, Hughes 1989:26).

Empirical Validity: Empirical Validity involves the comparison of the results of two separate and distinct tests (given at the same time) with the view to further validating the empirical data generated. In this case, the results of Part I and Part II can be compared and combined in order to provide more solid statistics on which to base decisions of course design (see 2.4.3, Heaton 1975:154, Lado 1961:324).

3.3 Factors that Affect Reliability and Validity

Certain intrinsic and extrinsic factors need to be considered to ensure adequate reliability and validity. Although they have so far been treated separately, reliability and validity will be discussed side by side in this section as they can often be affected by the same factors:

3.3.1 Lexical and Grammatical Items

Appropriate lexis and grammar are essential to the validity of the test. In
Part I, care should be taken to select minimal pairs of items that are of the same parts of speech (e.g. both nouns, both verbs etc.) otherwise the student could possibly deduce the correct answer from the grammatical or syntactic context of the test question. However, particularly with a group of mixed ability learners, there is a high possibility that a number of the lexical items included may not be within the knowledge of some testees, which may in turn mislead them to error. Curiously, some authors argue that these unknown items can hold potential advantages:

Valette (1967:57,60) claims that 'pure' discrimination tests are 'content-free': through recognition of the key phoneme alone the student should be able to select the correct answer without understanding the meanings of the words used, bringing the student to realise that the ability to differentiate between contrastive pairs may be crucial in understanding the language. Lado (1961:64) agrees that, in this respect, the inclusion of items "beyond the student's vocabulary yet within the phonemic contrasts and patterns" would therefore be valid. He also states that the students should be warned of the inclusion of such items before the test so that they would not be led astray.

Heaton (1975:13) notes however, that wild guessing reduces score reliability. In Part I, as with all two-choice tests, there is a 50% chance that the candidate could simply guess each answer correctly. The inclusion of unknown lexical items coupled with the inability to differentiate the correct answers from grammatical or syntactical contexts could encourage guesswork, reducing reliability. In contrast, for Part II, lexical items that are non-ambiguous, picturable, potentially within the student's active vocabulary and have only one current pronunciation should be selected.
otherwise it is unlikely that the testees would be able to create a sentence in which to produce them. In this respect, reliability is relatively higher.

### 3.3.2 Scoring of Results

The objective scoring system of Part I (see 2.4.1) would produce exactly the same results each time it was marked by any number of markers, provided all the markers used the same answer key. This system is therefore 100% reliable, also displaying very high inter-marker reliability.

The subjective scoring system of Part II (see 2.4.2) could present problems. The marker would have to apply the same judgement consistently and systematically to all the results. If more than one marker were involved they would have to negotiate the grading system together to maintain uniformity. It is worth noting, however, that the use of two markers would increase reliability even without the use of a grading system (Harrison 1983:112, Underhill 1987:88).

### 3.3.3 Use of Tape Recorders

The use of a recorded test in Part I carries high reliability as it ensures uniformity of test delivery every time it is used, particularly useful if the test group were too big to all take the test at once and needed to be divided into two groups, both requiring the same test. Attention must be paid to the quality of the recording, however. Poor quality will reduce validity.

The use of a tape recorder also has very practical implications when grading the results of Part II. The marker can replay the student's response
as many times as deemed necessary to properly assess production of any
particular speech sound, increasing scoring reliability (Lado 1961:47,

3.3.4 Learners' Sense of Familiarity

Reliability and validity can be enhanced if test candidates share a high and
equal amount of familiarity and ease with aspects of the testing process as
the opposite may hinder performance. The following factors affecting the
construction and administration of the test need to be taken into
consideration:

The testing environment: a quiet classroom with adequate acoustics,
comfortable temperature, suitable seating and lighting should be used.

The tester: the students' teacher would make the most valid tester as he or
she would be familiar to the students. The teacher being a real part of the
students' lives, tensions and artificialities would be lessened. Unfortu-
nately, it would be unlikely that a teacher familiar to the students would be
available, given the teaching situation described (see 1). The teacher
designated to the ensuing course would most likely administer the test,
perhaps even during the first lesson of the course.

The rubric: simple, clear examples, presented on the students' test sheets
and demonstrated by the tester, should illustrate the test before the process
actually begins. The wording and metalanguage of the instructions must
be considered within the students' lexicon as this can affect the validity of
the rubric itself (Hughes 1989:39, 106, Underhill 1987:40, Heaton 1975:58, 98,
160-161).
3.3.5 Psychological and Personal Factors

Poor health, fatigue, lack of interest or motivation, test-wisness, cognitive style (some such changes brought about by the time of day) all can affect an individual’s test performance and decrease reliability. Sex, age and learning background or experiences can also affect test scores but in the cases of reasonably homogeneous, monolingual classes of Japanese university students, all of very similar age and educational background (with the possible exceptions of study experiences abroad and individual degrees of contact with native English speakers) the risk to reliability is relatively low (Bachman 1990:160-161).

3.3.6 Time Restraints

Time restraints imposed on the students’ opportunities to answer each test question in both Parts I and II of the test can also affect reliability. If the time between the recorded questions in Part I and the time allowed for the student to produce a sentence in Part II can be considered reasonably representative of the timing between sentences in a ‘natural English conversation’, then reliability is less affected. However, such timing is extremely hard to quantify due to obvious variations among native speakers and conversational situations. The test environment of the classroom and the form of the test itself is also hardly representative of a ‘natural English conversation’.

In response to this problem, the timing of each test response should be kept short but not unnecessarily hurried as, when engaged in conversation with
a native English speaker, learners often find themselves with limited time to respond (delayed responses can disrupt the natural flow of a conversation or risk losing the attention of the listener). Therefore, during both parts of the test, only a few seconds should be allowed for students to select and tick the correct answer in Part I and produce a short sentence in Part II.

4. Conclusions

With test reliability and validity described as 'complimentary aspects' and 'two poles' (see 3) Hughes (1989:42) notes that "the tester has to balance gains in one against losses in the other". With this in mind, it is hoped that this test, divided into its two parts, reasonably satisfies these parallel aspects, creating a suitable balance. It is also hoped that the test measures and scores what it is supposed to test in a reasonable time, considering the testing situation and personnel available, and so may be regarded as practical, economical and easily administrable.

However, as with all language tests there are always areas that can benefit from improvement. Ideally, in order to more accurately elicit the learners' responses to (Part I) and production of (Part II) the test phonemes/speech sounds, the forms of both parts of the test should more greatly resemble natural English conversation. For example, for Part I, instead of a recording of unrelated sentences uttered in isolation, test speech sounds could be incorporated into paragraphs of sentences with a single context. On face value, this method would seem more immediately appropriate: aided by a common context, pure guesswork could be reduced, increasing
reliability, but Valette's claim that discrimination tests should be 'content free' (see 3.3.1) would not be addressed (assuming the test designer sympathises with this view). For Part II, instead of using picture stimuli to elicit the test speech sounds, the tester could attempt to create a spontaneous conversation with the learner, 'steering' the context of the conversation in ways to elicit the target speech sounds without the testee's knowledge of what was being tested. This method would again, seem more appropriate on face value but could decrease uniformity of test answers from student to student, in turn decreasing reliability (see 3.1).

In their proposed forms (see 2.3.2) the reliability of both Parts I and II could automatically be improved by doubling or tripling the number of test items for each phoneme/speech sound, confirming the diagnosis of areas of weakness and reducing the influence of guesswork in Part I. Although sound, the practicality of this would largely depend on the time available for testing the students as a class and as individuals (in Part II). In cases of very large groups and very limited time for testing, the teacher may have to resort to only using Part I of the test. Also, whether the tester felt that this would affect the spontaneity of students' answers due to growing test-wiseness or other personal or psychological factors (see 3.3.5) that can have detrimental effects on students' performance.

The two-grade distinction between 'satisfactory' and 'poor', used in the marking system for Part II (see 2.4.2) is arguably inappropriate to single phonemes as it does not accurately highlight students' individual weaknesses in articulation. Such weaknesses could be more accurately diagnosed through comparisons of students' recorded test results to samples of recordings of native English speakers, or better still by comparison of wave
forms of the students' speech on an oscilloscope or spectrograph to that of native speakers. Ideally, the tester would have received some phonetic training with which to administer such procedures effectively. The test results would be able to more effectively influence subsequent pronunciation teaching as well as general test design.

Unfortunately, many testers may not have the necessary training, time or equipment to conduct such an analysis. Although less accurate, the method of grading test results outlined in 2.4.2 would at least provide quick and easy comparison between the results generated in Part I and reasonably inform the teacher as to which speech sounds to make priority items during the ensuing course.

In conclusion, although imperfect, the proposed test could be quickly and easily administered to large groups of the learners described, generating enough information for the short-term, one-semester teacher to use to shape the subsequent course.

Bibliography


Davies, Alan (1990), Principles of Language Testing. Basil Blackwell Ltd.


Heaton, J. B. (1975), Writing English Language Tests. Longman.

University Press.


Lado, Robert (1961), Language Testing. Longmans, Green and Co. Ltd.


