RELIABILITY AND VALIDITY OF A TEST AND ITS PROCEDURE CONDUCTED AT A JAPANESE HIGH SCHOOL

Paul Bela Nadasdy*

Abstract

This paper analyses the validity and reliability of an English listening test being used at a private Japanese high school. Through an analysis of the test results, an attempt was made to make salient the qualities and deficiencies of the test and its procedure.

The test's reliability was analysed using a split-half method measuring the coefficient of internal consistency. The split-test's coefficient results suggested that there was a certain amount of unreliability between the two halves of the test. Although the reliability was below an acceptable level, calculations using the Spearman-Brown formula suggested the possibility of higher coefficiency. Regarding construct, content, criterion-related, and face validity the test appeared valid.

Key words : TESTING, VALIDITY, RELIABILITY, HIGH SCHOOL

1. Introduction

Though some have argued whether testing is actually necessary at all, it is generally agreed that it is the most practical way to monitor and systematically rank students. And as tests remain the most popular way to grade students fairly, the quality of their production would seem vital.

For test efficiency, validity and reliability need to be present. And as these two conditions are important for the effectiveness of testing, it is generally accepted that we can achieve a precise evaluation of our students if they are both consistent. Unsurprisingly, however, the variables that exist in measuring both reliability and validity in tests at times produce a range of results.

This paper starts with an analysis of testing in general and of how the examination of validity and reliability is used as a means of quality control in test production. This is followed by an analysis of a listening test that is being used in a high school in Japan. Quantitative and qualitative results are analysed to ascertain whether it is reliable and valid, and this followed by a evaluation of its overall effectiveness.

2. Review of literature

Analysts who have made important contributions within the realm of testing include

*Paul Bela Nadasdy［情報文化学科］

Oller (1979), Hughes (1989), Bachman (1990), Spolsky (1985), Messick (1996), Fulcher (1997), Cohen et al. (2000), and Chapelle (1999, 2003). In defining testing and its usefulness Bachman states that "language tests are indirect indicators of the underlying traits in which we are interested" (1990 : 33). Davies (1990), Hughes (1989), and Baker (1989) refer to tests in the way that they help us to acquire information, act as a procedure for problem solving, and act as a decision making procedure (Owen 1997 : 2). Owen also offers an endorsement of testing in that instructors need to monitor student progress independently, which opposes the possibility of inaccurate and biased self-assessment (1997 : 5).

Owen further defines possible motivations for tests in language learning explaining that they assist in ranking students, assist in gauging whether students are able to cope with certain language forms, help us to observe whether learning has been achieved, give useful information relating to forecasting future developments in student performance, and help us to refine what we are teaching and testing. Furthermore, testing can also contribute to establishing whether certain entities are effective such as teachers, schools and teaching methods in comparing them against one another. Among these positive endorsements, Owen also suggests that tests act as a means of control and motivation of our students. However, some commentators draw our attention to the negative reputation that tests have within the teaching community. For example, Hughes (2003) refers to the "mistrust" educators have of tests and testing in general.

3. Validity in testing

Two areas should be considered when discussing validity in testing :

1. Consider how closely the test performance resembles the performance we expect outside the test.
2. Consider to what extent evidence of knowledge about the language can be taken as evidence of proficiency.

(Owen 1997 : 13)

Referring to the importance of validity in tests, Cohen et al. (2000) state that effective research is impossible or even "worthless" without the presence of validity (2000 : 105), though they do recommended against aiming for absolute validity. Instead they define the search for validity as being one of minimizing invalidity, maximizing validity, and therefore using measurement in validity as a matter of degree rather than a pursuit of perfection (2000 : 105). Owen (1997) citing Baker (1989) also considers the accuracy and proficiency of testing and how we evaluate individuals :

It is quite useful for understanding tendencies in testing, but...it seems less easy actually to allocate particular tests to one cell rather than another, and...it is not easy to separate

knowledge of system as a counterpoint to performance from knowledge of a system as indirect evidence of proficiency.

(Owen, 1997 : 17)

3.1 Construct, content, criterion-based, and face validity

Several categories exist for validity. The following four categories are described by Hughes (1989) and Bachman (1990), these being construct validity, content validity (included within this are internal and external validity), criterion-based validity, and face validity.

3.1.1 Construct validity

Construct validity is concerned with the level of accuracy a construct within a test is believed to measure (Brown 1994 : 256 ; Bachman & Palmer 1996) and, particularly in ethnographic research, "must demonstrate that the categories that the researchers are using are meaningful to the participants themselves" (Cohen et al 2000 : 110).

3.1.2 Content validity

Content validity is concerned with the degree to which the components of a test relate to the real-life situation they are attempting to replicate (Hughes 1989 : 22 ; Bachman 1990 : 306) and is relevant to the degree to which it proportionately represents. Within the domain of content validity are internal validity and external validity. These refer to relationships between independent and dependent variables when experiments are conducted. External validity occurs when our findings can be related to the general populous, whereas internal validity is related to the elimination of difficult variables within studies.

3.1.3 Criterion-related validity

Criterion-related validity "(relates) the results of one particular instrument to another external criterion" (Cohen et al. 2000 : 111). It contains two primary forms, these being predictive and concurrent validity. Concerning predictive validity, if results from two separate but related experiments or tests produce similar results the original examination is said to have strong predictive validity. Concurrent validity is similar but it is not necessary to have been measured over a span of time and can be "demonstrated simultaneously with another instrument" (2000 : 112).

3.1.4 Face validity

This term relates to what degree a test is perceived to be doing what it is supposed to. In general, face validity in testing describes the look of the test as opposed to whether the test is proved to work or not.

3.2 Messick's framework of unitary validity

Messick's (1989) framework of unitary validity differs from the previous view which

identifies exclusively content validity, face validity, construct validity, and criterion-related validity as its main elements. Messick considers these sole elements to be inadequate and stresses the need for further consideration of complementary facets of validity, and in particular the examination of scores and construct validity assessment as its key features. Six aspects of validation included in Messick's paradigm provide "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (Messick 1989 : 13 cited in Bachman 1990 : 236).

These elements are judgmental/logical analysis, which is concerned with content relevance, correlation analyses, which utilizes quantitative analyses in interpreting test scores to gather evidence in support of the test scores, analyses of process, which involves the investigating of test taking, analyses of group difference and change over time, which examines to what extent score properties generalize across population groups, manipulation of tests and test conditions, which is concerned with gathering knowledge about how test intervention affects test scores, and test consequences, which examines elements that affect testing including washback, consequences of score interpretation, and bias in scoring (Bachman 1990 ; Messick 1996).

3.3 Testing outcomes

Considering the above framework defining validity in testing, we need to consider the importance of determining what is appropriate for our students and teaching situations as well as on a larger scale. The importance of analysis in low-stakes testing could be significant if one considers how data can be collected from the source and used productively. Regarding Chapelle's (2003) reference to Shepard (1993) in that the primary focuses are testing outcomes and that "a test's use should serve as a guide to validation" (2003 : 412), suggests we are in need of a point from where to start our validation analysis from. Chapelle also cites that "as a validation argument is 'an argument' rather than a 'thumbs up/thumbs down' verdict" (Cronbach cited in Chapelle 2003), we start to focus on something that we can generally agree is an important outcome — the result.

4. Reliability in testing

Reliability relates to the generalisability, consistency, and stability of a test. Following on from test validity Hughes points out that "if a test is not reliable, it cannot be valid" (2003 : 34). Hughes continues that "to be valid a test must provide consistently accurate measurements" (2003 : 50) Therefore it would seem that the higher amount of similarity there is between tests, the more reliable they would appear to be (Hughes : 1989). However, Bachman (1990) argues that although the similarity case is relevant, other factors concerning what we are measuring will affect test reliability. Factors including test participants' personal characteristics i.e. age, gender, and factors regarding the test environment and condition of

the participants can contribute to whether or not a test is effectively reliable (1990 : 164).

Investigating reliability can also be approached by analyzing a test candidate's Classical True Score (CTS). According to Bachman (1990 : 167), concerning CTS, if it was possible for a test candidate to take the same test in an unaffected environment several times, it is conceived that the eventual mean score would provide a total that would closely equate to the participants true score. In using CTS one can calculate reliability and especially reliability coefficients in three areas — internal consistency, test score validity over a period of time, and in comparing forms of tests (1990 : 172-182). What is ascertained from the CTS is no doubt important. However, the results are still in theoretical realms and may not take into account variables that could be established via empirical investigations.

In considering that even in strict testing conditions conducted at different times human changeability is unavoidable and the same test conducted twice in similar conditions will provide conflicting results. With regards to this one may wonder how possible it would be to test reliability. However, taking into consideration the 'reliability coefficient' which helps to compare the reliability of test scores, we may start to get closer to determining test reliability. One can aim for similar scores that fall within an acceptable range and observe a mean average that signifies reliability (the reliability coefficient).

Terms relating to reliability can be defined in the following ways. Inter-rater reliability is concerned with how scores from various sources are balanced and importantly to what degree markers scores are showing equality (Nunan 1992 : 14-15). Test-retest reliability gives an indication as to how a test consistently measures individual performances of students that are tested across various testing organizations (Underhill, 1987 : 9). A further simplified definition is offered by Nunan and Weir and Roberts stating that inter-rater reliability is the degree to which the scores from two or more markers agree (Nunan 1992 : 14-15 ; Weir and Roberts 1994 : 172). Examples of methods estimating reliability include test-retest reliability, internal consistency reliability, and parallel-test reliability. These methods each have their own ways of examining the source of error in testing.

5. Procedures to ensure validity and reliability

5.1 Ensuring validity

Hughes states that the concept of test validity can seem uncomplicated but on closer inspection can appear highly complex (2003 : 34). Some experts say that "one might suppose that ultimately there is no means of knowing whether a test is valid or not." (Owen 1997 : 13) One certainty is that it is possible to describe and assess test validity in various ways. Initially, one could attest that the most important description is based around test effectiveness. Hughes (2003) points out the basis for a simple criterion for test quality and offers evidence for showing relevance of certain descriptions that may help to rectify difficulties in

language testing. Firstly, he states specifically that a test should simply "... (measure) accurately what it is intended to measure" (2003 : 26) to assure us of its validity.

Though this may appear relatively simple in terms of straightforward testing, several definitions of what we expect our students to achieve can overcomplicate what we are attempting to measure. To assist in simplifying ambiguous "theoretical constructs" such as fluency in speaking, reading ability etc. certain descriptions of validity can be considered including construct validity, content validity, and criterion-related validity. The following considers these variants. With content validity, Hughes points out that if the test has positive content validity it is more likely to accurately test what is required, and thus leads to construct validity. He states that "the greater a tests content validity, the more likely it is to be an accurate measure of what it is supposed to measure" (2003 : 27). Importantly, when creating tests, specifications have to be established at an early stage referring to what is required from the tests participants. These specifications should be areas that are considered to be of maximum benefit when defining that which is to be measured and achieved through the testing. Hughes purports though that "too often the content of tests is determined by what is easy to test rather than what is important to test" (2003 : 27). Therefore it is important to be clear about what is required. Criterion-related validity provides assessment from different perspectives and presents an opportunity to compare qualitative score analysis against quantitative independent judgments of test participants' abilities. Hughes states that all of these "have a part to play in the development of a test" (2003 : 30).

Hughes also draws our attention to how scoring is important when judging the validity of tests and how testers and test designers must "make sure that the scoring of responses relates directly to what is being tested" (2003 : 34). Accurate scoring of responses would seem imperative if correct measurement is to be assured. Being clear as to what is required as a response e.g. clear responses of pronunciation on speaking tests should not be confused with hesitation or intonation issues, validity may then be more achievable and measurements more accurate and relevant.

5.2 Ensuring reliability

According to Hughes there are several ways to ensure reliability. These include gathering information about the test candidate by adding extra and more detailed questions, tasks, and examples to tests, balancing the difficulty of questions so they do not "discriminate between weaker and stronger students", focusing and restricting questions that may allow for too much elaboration, avoiding ambiguous questions and items, being clear with instructions for tasks, presenting tests clearly to avoid confusion, practicing the test format with students so that they are familiar and prepared for the actual test, encouraging consistency across administrations on large scale testing, using items that utilize objective scoring i.e. providing part of an answer for a test taker to complete rather than eliciting an entire sentence as an

answer, restricting the freedom afforded to candidates in terms of the comparisons made between them, providing clear and detailed score keys, helping testers and scorers by training them at an early stage and conferring with test designers and testers about how responses are to be scored before scoring commences, having students represented by numbers rather than personal details to restrict any possible bias occurring, and using, if possible, independent scorers to evaluate objectively eliminate discrepancies (1989 : 44-50).

Though the variable in human errors in testing between testers and candidates are significant, these items seem to at the very least work towards creating better reliability. It would certainly seem of benefit to have practical experience of teaching and testing enabling researchers a firsthand experience of what may be required throughout the entire process of test organization.

6. Method

6.1 Listening Test

The test selected for this analysis is designed for testing the listening ability of 1st grade students who are in their second term at a senior high school in Japan. Preparation for the test is conducted over a period of three weeks prior to the actual test which is given in the fourth week of each month respectively.

The test is one of several listening tests conducted each term and is administered over the period of two weeks for approximately five hundred first grade students. Ten native speaking English teachers design, administer, and mark the test. Totals are added to the students' final yearly grade and are important for graduation..

The test conditions require students to listen to a 20 minute recording of monologues and dialogues relating to a syllabus item designated for that particular month. The test chosen for this study consists of four sections relating to 'favourites', 'possessives', 'numbers', 'jobs', and 'personal information'

6.2 Split-half analysis

With a view to narrowing down the variables that might affect consistency in measuring reliability within the research, a singly-administrated split-half method (Hughes 1989 : 40) in which the "coefficient of internal consistency" (1989 : 40) can purportedly be measured was utilized. The test was designed so it could be separated into relatively equal parts in order to collect two separate scores following a single session. One class of thirty upper-intermediate test participants was selected for the analysis.

7. Results

Lado (1961) cited in Hughes (1989 : 39) suggests that a good listening test should fall in the

range of 0.80-0.89 reliability coefficient. The split-test's coefficient results (see table 1 (1)) suggest that there is a certain amount of unreliability between the two halves of the test. With the coefficient score of .36 there are obvious underlying problems.

However, though the test scores identically between part 1/2 to 3/4 the test does vary in minor degrees in contents which could have caused discrepancies within the consistency between the two sections (see Appendix 5). In order to establish whether reliability was affected by task order and/or task groupings, the test was analysed in different ways. The reliability coefficient was analysed after collecting odd and even scores, from calculating various test task groups together, and by calculating split tasks which were connected to each equivalent on the opposite part of the test (see table 2).

The reliability coefficient results were as follows:

Table 1

Calculation type	Coefficient
(1) Questions 1-25/26-50	0.36
(2) Every other question	0.70
(3) Tasks 1/4 & 2/3	0.77
(4) First/second halves of tasks	0.73

Though the original measurement of reliability was relatively low, it can be observed that by varying the way in which the coefficient is calculated higher scores of coefficients can be achieved. This suggests that there may be a certain amount of reliability in the test. Dividing the total scores by the four types of analysis equates to the following sum:

Calculations (4)÷Coefficient total (2.56) = **0.64**

Applying the Spearman-Brown formula (Reliability = 2r÷1+r), the possibility of higher coefficiency was investigated. The results were as follows:

Table 2

Calculation type	Coefficient	Spearman-Brown
Questions 1-25/26-50	0.36	0.53
Every other question	0.70	0.82
Tasks 1/4 & 2/3	0.77	0.87
First halves of tasks/second halves of tasks	0.73	0.84

The averaged coefficient of **0.64** was then calculated using the Spearman-Brown model giving the final internal consistency score :

$(0.64 \times 2 = 1.28/0.64 + 1 = 1.64) = 1.28 \div 1.64 =$ **0.78**

Considering Lado's (1961) estimates of 0.80-0.89, this final score falls just below a satisfactory level of reliability.

8. Test evaluation

In design, the four sections of the test mirrored each other to compensate for the subsequent split-half analysis. Each task was evaluated using the model described by Hughes's (1989) construct validity, criterion-related validity, content validity, and face validity model. Following an analysis into the test's validity further investigations were made to establish its level of reliability.

In terms of construct validity it is important for the test items to measure what they are supposed to and to be meaningful to the test participants. Part 1 (see Appendix 5) consists of four questions relating to favourite people and items. Taking into consideration the cultural differences between Japanese and western students it would be difficult to guarantee that the items are completely meaningful. However, as the test does contain popular figures and well know items these will at least have some appeal at a basic level, and in a small scale test without independent evaluation seems relevant. Parts 2 and 3 (see Appendix 6/7) are both related to numbers. How useful this construct will be to student may be ascertained when applied practically.

As well as evaluate students' listening abilities and contributing to the students final grade, the test items are designed to prepare students for a post-course homestay in the UK. The recordings are all in British English and contain natural speed and rhythm. If the preparation is effective and if the students go on to recognize or use these items, this appears to validate the construct. Part 3 (see Appendix 7) consists of information relating to nationality, profession, and city of residence. As this test is part of an ongoing course dedicated to helping students retain language items, this repeated strategy seems adequate in its inclusion.

Regarding content, criterion-related, and face validity the items in the test are recordings extracted from the students' coursebook and are clear recordings of speech mainly in British accents. Internally the contents seem to appear valid in that they attempt to replicate real-life situations. However, the unnatural delivery suited for second language students challenges whether this is completely content valid. In terms of criterion-related validity, there is only one instrument of measurement in this study, so it would therefore be difficult

to make comparisons with other examples. Comparing the two halves of the split-test analysis there seems to be discrepancies (see Table 1). The variation in scores in some respects proves that there may be some weakness in the use of some of the contents. In terms of face validity, the test has the appearance that it will work well as a listening test with ample examples that are easy enough to follow, simple legible instructions, and coverage of a sufficient range of language items.

In terms of reliability, the conditions were quite varied during the test's design, administration, and scoring. As there were several teachers and designers involved in the process, it became apparent that it would be difficult to prove exactly how reliable and consistent the test was. It can be observed from the tables (see Appendix 1-4), however, that the participating students scored very well on the test. Comparing these mark to other classes, they generally scored higher in most cases. As the class consisted of upper-intermediate level students, the scores achieved were close to what was expected ; the scores being similar to the students' regular grades. As the testing conditions and marking were conducted by one individual, this reduced interference by outside influences. Though students were required to write their details on the front page (see Appendix 5) the scoring was unbiased and consistent. In conclusion, the test seemed to achieve what it was meant to. It tested items that were meaningful to the students, covered the school syllabus, achieved an expectation relating to scores, reinforced language items and tested students' recognition of language in context, and worked well in general as an auditory test.

9. Conclusion

As mentioned in the introduction of this paper, high stakes test are causing further demands to be met by test designers in creating tests that accurately measure what they are supposed to. As Hughes states, designers of tests must try to "make their tests as valid as possible" (1989 : 34). Details regarding the validity and reliability of tests should be made available so there can be careful observation of how and what tests are measuring. If the general consensus about a test is good, it can be considered as a benchmark for designers to work from. Though, as mentioned in the background, as the pursuit of perfection is perhaps ultimately unproductive, we can instead strive to encourage communication across administrators, designers, and teachers to improve what we are ideally working towards — more validity and reliability in tests and less invalidity and unreliability (Cohen et al. 2000).

Referring again to what tests are intending to measure, we can strive towards creating test items that truly elicit meaningful, appropriate, and measurable language forms from learners in order to evaluate ability. It would seem the problem is in defining what exactly to look for in proficiency. In terms of establishing this, it could be said that the closer one is to the source e.g. the classroom, students, test design, the better chance there would be of achieving this.

Appendix 1

11.1 Split half test : 1-25/26-50

Student	Score 1 (25)	Score 2 (25)	50pts = 100%	Variability
1	24	25	49 (98%)	1 (2%)
2	20	22	42 (84%)	2 (4%)
3	20	24	44 (88%)	4 (8%)
4	25	25	50 (100%)	0 (0%)
5	24	25	49 (98%)	1 (2%)
6	20	24	44 (88%)	4 (8%)
7	20	25	45 (90%)	5 (10%)
8	24	25	49 (98%)	1 (2%)
9	25	20	45 (90%)	5 (10%)
10	20	22	42 (84%)	2 (4%)
11	24	25	49 (98%)	1 (2%)
12	24	25	49 (98%)	1 (2%)
13	25	25	50 (100%)	0 (0%)
14	18	20	38 (76%)	2 (4%)
15	15	20	35 (70%)	5 (10%)
16	20	25	45 (90%)	5 (10%)
17	24	24	48 (96%)	0 (0%)
18	25	25	50 (100%)	0 (0%)
19	24	20	44 (88%)	4 (8%)
20	20	20	40 (80%)	0 (0%)
21	24	24	48 (96%)	0 (0%)
22	24	25	49 (98%)	1 (2%)
23	25	25	50 (100%)	0 (0%)
24	24	25	49 (98%)	1 (2%)
25	24	20	44 (88%)	4 (8%)
26	24	22	46 (92%)	2 (4%)
27	20	18	38 (76%)	2 (4%)
28	24	20	44 (88%)	4 (8%)
29	20	24	44 (88%)	4 (8%)
30	20	20	40 (80%)	0 (0%)
	Co-efficient : 0.361618		100% Equality = 8 (times) 26.6%	

Appendix 2

11.2 Split half test : every other question

Student	Score 1 (25)	Score 2 (25)	50pts = 100%	Variability
1	24	25	49 (98%)	1 (2%)
2	21	21	42 (84%)	0 (0%)
3	22	22	44 (88%)	0 (0%)
4	25	25	50 (100%)	0 (0%)
5	24	25	49 (98%)	1 (2%)
6	22	22	44 (88%)	0 (0%)
7	21	24	45 (90%)	3 (6%)
8	25	24	49 (98%)	1 (2%)
9	21	24	45 (90%)	3 (6%)
10	22	20	42 (84%)	2 (4%)
11	25	24	49 (98%)	1 (2%)
12	24	25	49 (98%)	1 (2%)
13	25	25	50 (100%)	0 (0%)
14	19	19	38 (76%)	0 (0%)
15	16	21	35 (70%)	5 (10%)
16	22	23	45 (90%)	1 (2%)
17	24	24	48 (96%)	0 (0%)
18	25	25	50 (100%)	0 (0%)
19	22	22	44 (88%)	0 (0%)
20	20	20	40 (80%)	0 (0%)
21	24	24	48 (96%)	0 (0%)
22	25	24	49 (98%)	1 (2%)
23	25	25	50 (100%)	0 (0%)
24	24	25	49 (98%)	1 (2%)
25	23	21	44 (88%)	2 (4%)
26	24	22	46 (92%)	2 (4%)
27	20	18	38 (76%)	2 (4%)
28	22	22	44 (88%)	0 (0%)
29	24	20	44 (88%)	4 (8%)
30	20	20	40 (80%)	0 (0%)
	Co-efficient : 0.696771		100% Equality = 14 times 46.6%	

Appendix 3
11.2 Split half test : part 1/4 and part 2/3

Student	Score 1 (25)	Score 2 (25)	50pts = 100%	Variability
1	24	25	49 (98%)	1 (1%)
2	20	22	42 (84%)	2 (4%)
3	21	23	44 (88%)	2 (4%)
4	25	25	50 (100%)	0 (0%)
5	24	25	49 (98%)	1 (2%)
6	21	23	44 (88%)	2 (4%)
7	24	21	45 (90%)	3 (6%)
8	25	24	49 (98%)	1 (2%)
9	22	23	45 (90%)	1 (2%)
10	21	21	42 (84%)	0 (0%)
11	24	25	49 (98%)	1 (2%)
12	24	25	49 (98%)	1 (2%)
13	25	25	50 (100%)	0 (0%)
14	20	18	38 (76%)	2 (4%)
15	17	18	35 (70%)	2 (4%)
16	21	24	45 (90%)	3 (6%)
17	23	25	48 (96%)	2 (4%)
18	25	25	50 (100%)	0 (0%)
19	22	22	44 (88%)	0 (0%)
20	18	22	40 (80%)	2 (4%)
21	23	25	48 (96%)	2 (4%)
22	24	25	49 (98%)	1 (2%)
23	25	25	50 (100%)	0 (0%)
24	24	25	49 (98%)	1 (2%)
25	22	22	44 (88%)	0 (0%)
26	22	24	46 (92%)	2 (4%)
27	19	19	38 (76%)	0 (0%)
28	22	22	44 (88%)	0 (0%)
29	22	22	44 (88%)	0 (0%)
30	18	22	40 (80%)	2 (4%)
	Co-efficient : 0.768211		100% Equality = 10 times 33.3%	

Appendix 4

11.3 Split half test : first halves of tasks/second halves of tasks

Student	Score 1 (25)	Score 2 (25)	50pts = 100%	Variability
1	25	24	49 (98%)	1 (2%)
2	22	20	42 (84%)	2 (4%)
3	22	22	44 (88%)	0 (0%)
4	25	25	50 (100%)	0 (0%)
5	25	24	49 (98%)	1 (2%)
6	21	23	44 (88%)	2 (4%)
7	21	24	45 (90%)	3 (6%)
8	24	25	49 (98%)	1 (2%)
9	22	23	45 (90%)	1 (2%)
10	20	22	42 (84%)	2 (4%)
11	24	25	49 (98%)	1 (2%)
12	24	25	49 (98%)	1 (2%)
13	25	25	50 (100%)	0 (0%)
14	19	19	38 (76%)	0 (0%)
15	16	19	35 (70%)	3 (6%)
16	22	23	45 (90%)	1 (2%)
17	24	24	48 (96%)	0 (0%)
18	25	25	50 (100%)	0 (0%)
19	24	20	44 (88%)	4 (8%)
20	20	20	40 (80%)	0 (0%)
21	24	24	48 (96%)	0 (0%)
22	25	24	49 (98%)	1 (2%)
23	25	25	50 (100%)	0 (0%)
24	24	25	49 (98%)	1 (2%)
25	22	22	44 (88%)	0 (0%)
26	24	22	46 (92%)	2 (4%)
27	20	18	38 (76%)	2 (4%)
28	22	22	44 (88%)	0 (4%)
29	21	23	44 (88%)	2 (4%)
30	20	20	40 (80%)	0 (0%)
	Co-efficient : 0.728156		100% Equality = 12 times 40%	

Appendix 5

ENGLISH MONTHLY TEST: GRADE 1
TERM 1 TEST 1 – SEPTEMBER 200
LISTENING TEST

NAME:

STUDENT NUMBER:

MARK: ____ / 50

PERCENTAGE: .....%

Part 1

Circle each person's favourite things. You will hear everything twice.

1) Samantha Jones (example)

Favourite singer:	* Beyonce	* Madonna	* Janet Jackson	* Michael Jackson
Favourite actor:	* Brad Pitt	* Will Smith	* Ben Affleck	* Jonny Depp
Favourite group:	* Beatles	* Green Day	* Beastie Boys	* Backstreet Boys
Favourite team:	* LA Lakers	* Chicago Bulls	* Miami Heat	* New York Knicks

2) Tony Soprano

Favourite food:	* pizza	* steak	* curry	* spaghetti
Favourite car:	* Toyota	* Ford	* BMW	* Ferrarri
Favourite singer:	* Madonna	* Kylie	* Janet Jackson	* Pavarotti
Favourite sport:	* football	* basketball	* baseball	* boxing

3) Frank Burnside

Favourite food:	* hamburgers	* pizza	* fish	* fish and chips
Favourite drink:	* milk	* tea	* whisky	* water
Favourite team:	* Chealsea	* West Ham	* Manchester Utd	* Arsenal
Favourite group:	* East 17	* West 16	* North 15	* South 14

4) Tom Glutton

Favourite food:	* ice cream	* hamburgers	* fried chicken	* pizza
Favourite drink:	* cola	* coffee	* tea	* orange juice
Favourite actress:	* Cameron Diaz	* Nicole Kidman	* Jennifer Anniston	* Jennifer Lopez
Favourite country:	* U.S.A	* Mexico	* Yugoslavia	* Ukraine

Part 2

You will hear nine telephone numbers. Tick the numbers you hear.

(Example) 1 ☐ 313557 ☑ 313597	4 ☐ 0509 23092 ☑ 0519 23092	7 ☐ 068 90 789 ☑ 068 91 789
2 ☑ 743678 ☐ 743670	5 ☑ 0457 64332 ☐ 0457 64323	8 ☐ 335278 ☐ 335279 ☑ 339279
3 ☐ 01 800 7689 ☐ 01 808 7680 ☑ 01 808 7688	6 ☐ 041 914 5389 ☑ 041 904 5308 ☐ 041 940 5388	9 ☑ 0425 5781 ☐ 0425 5718

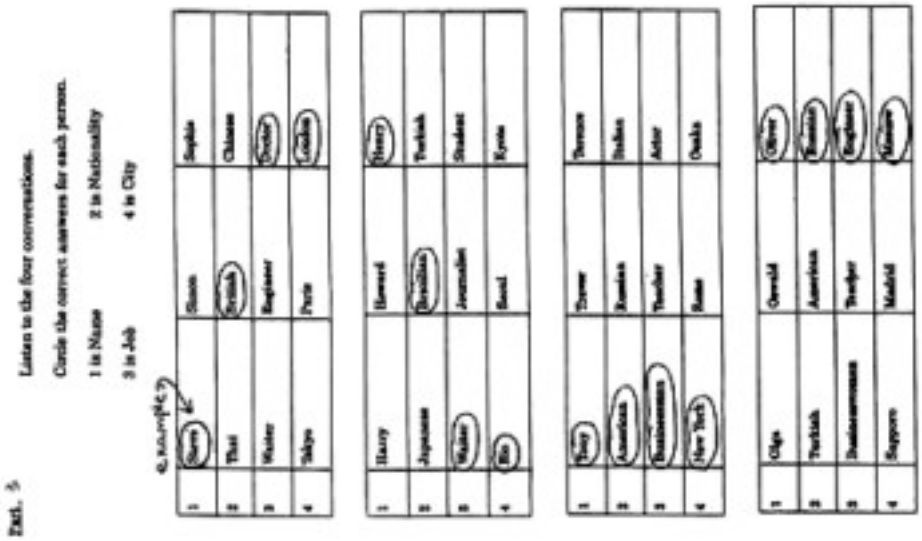
Listen to people asking for telephone numbers.
Write down the correct numbers.

(example) Odeon Cinema..... 091 747 6443

1. Pizza La..... 021 930 2738
2. Waseda University..... 061 439 4576
3. Kawagoe City Office..... 031 388 0542
4. British Airways..... 031 897 4567

Appendix 7


Part 4
Listen to the numbers and write the letter
C eighty
D ninety-two
thirty-five
sixty-eight
H fourteen
E forty
I eighteen
B seventeen
K seventy-nine
example A twenty-one
fifteen
G eighty-three
fifty-two
twenty-five
J thirteen
F sixty-one


Part 3
Listen to the four conversations.
Circle the correct answers for each person.
1 is Name 2 is Nationality
3 is Job 4 is City
example
Steve Sam Sophie
1 Steve British Chinese
2 Thai Waiter Doctor
3 Waiter Engineer London
4 Tokyo Paris
Harry Howard Harry
1 Japanese Brazilian Turkish
2 Waiter Journalist Student
3 Rio Basel Kyoto
4
Tony Trevor Theresa
1 American Russian Italian
2 Businessman Teacher Actor
3 New York Rome Osaka
4
Olga Oswald Olivia
1 Turkish American Russian
2 Businesswoman Teacher Engineer
3 Sapporo Madrid Moscow
4

References

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

Bachman, L. & A. Palmer (1996). *Language Testing in Practice*. Oxford University Press.

Baker, D. (1989). *Language testing : a critical survey and practical guide*. Edward Arnold.

Brown, H.D. (1994). *Principles of Language Learning and Teaching* (3rd. ed.) Prentice Hall.

Chapelle, C.A. (1999). Validation in language assessment. *Annual Review of Applied Linguistics* 19, 254-72.

Chapelle, C.A., Jamieson, J. & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, 20(4), 409-439.

Cohen, L., Manion, L. & Morrison, K. (2000). *Research Methods in Education*. Routledge/ Falmer.

Davies, A. (1990). *Principles of language testing*. Oxford, UK ; Cambridge, Mass., USA : B. Blackwell, 1990.

Fulcher, G. (1997). "Assssing Writing." In Fulcher, G. (ed.) *Writing in the English Language Classroom*. Hemel Hempstead : Prentice Hall Europe.

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

Lado, R. (1961). *Language Testing*. London : Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York : Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*. ETS. Princeton.

Nunan, D. (1992). *Research methods in language learning*. Cambridge University Press.

Oller, J. (1979). *Language tests at school : A pragmatic approach*. London : Longman.

Owen, C. (1997). *Testing*. Birmingham : The Centre for English Language Studies.

Read, J. & Chapelle, C.A. (2001). A framework for second language vocabulary assessment. *Language Testing*, 18(1), 1-32.

Shepard, L.A. (1993). "Evaluating Test Validity." In L. Darling-Hammon (Ed.), Review of Research in Education, Vol. 19. Washington, DC : AERA.

Spolsky, B. (1985). Formulating a theory of second language learning. *Studies in Second Language Acquisition*, 7, 269-288.

Underhill, N. (1987). *Testing Spoken Language : A handbook of oral testing techniques*. Cambridge University Press.

Weir, C. & J. Roberts (1994). *Evaluation in ELT*. Blackwell Publishing.