Reliability and Validity of a Test and its Procedure Conducted at a Japanese High School

Paul Nadasdy*

Abstract

An English listening ability test distributed at a Japanese high school was analysed in order to test its reliability and validity. A split-half analysis was used to test the "coefficient of internal consistency" (Hughes 1989) and the reliability and validity was further analysed utilising the Spearman-Brown formula in order to identify higher coefficiency. Results showed that, though improvements could be made to its overall content, the test was both relaible and vaild within the parameters of the models used and that both construct validity and face validity were sound.

Keywords : testing, reliability, validity, listening tests

Introduction

The significance of testing and of accuracy in what tests measure is currently under intense scrutiny. The importance of tests, in particular 'high stakes' testing, ensures that the quality of test production is vital. Though some have argued whether testing is actually necessary at all, it is generally agreed that tests are important to the monitoring and systematic ranking of students. Owen offers an endorsement of testing in that instructors need to monitor student progress, and "…since self-assessment is arguably open to abuse, it is generally and less subjective i.e. a valid test, is required for many social purposes" (Owen 1997: 5).

Within the domain of educational research two categories which have contributed to recent qualitative and quantitative research into the effectiveness of testing are salient. These are the categories of validity and reliability. One may attest that by paying close attention to the reliability and validity of tests we can achieve effective evaluation of our students. Unsurprisingly, however, the variables that exist in measuring both reliability and validity in tests at times produce a range of results.

The paper, limited as it is as a piece of individual research, intends to analyse the validity and reliability of an auditory comprehension test that is used in an educational setting in Japan. Through an analysis of the test results, an attempt will be made to make salient its qualities and its deficiencies and attempt to point out how it could be improved. It is hoped that the suggested improvements could then be applied to make the test more consistent, and the results and conclusions applied to wider context.

Firstly, this paper will examine the

＊工学部英語系列講師　Lecturer, Department of English Language, School of Engineering

various views of analysts concerned with definingthe purpose of testing and offer a breakdown of the various manifestations of validity and reliability in testing. This is followed by an analysis of a test that is being used in a professional teaching situation in Japan. Quantitative results are analysed and qualitative analysis is applied to the research. Finally there is an analysis on how the results can affect teaching in general.

Background

Commentators who have made important contributions within the analysis of testing include Oller (1979), Hughes (1989), Bachman (1981, 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 respectively (Owen 1997:2). Owen points out, however, the difficulties that surface in defining what sort of problem and what sort of decision that is needed to be made (1997:2). Though written mainly from a teacher perspective, Owen further defines possible motivations for tests in 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. Owen also suggests that tests act as a means of control and motivation of our students. These views, one could suggest, come from practical experience of analysing student needs rather than from a purely analytical basis and one which is primarily concerned with generalisations.

Some commentators on testing draw our attention to the negative reputation that tests have within the teaching community. Hughes (2003) refers to the “mistrust” educators have of tests and testing in general. The quality of testing may have a lot to do with the level of experience of test designers, but maybe more importantly problems arise when taking into consideration what the test is in fact intended to measure. Its failure to achieve what it sets out to do, it can be said, may be detrimental to the teaching and learning environment. This supposed failure to test what is being taught may be related to how tests are designed and administered. Tests created on a large scale may have less to do with what is taught in the classroom than what is actually considered beneficial based on education authority standards. In the subsequent analysis it can be observed how a small-scale test designed at source may contribute to assuring validity and reliability.

The failure of tests to measure true abilities of students appears to be relatively common. Content and test techniques have a lot to do with how tests may result in inaccuracies. Another concern is how consistent a test can be. If a test measures consistently it may appear we can be confident of its reliability. Human inconsistency will be a factor in how scores vary but this differential should not be markedly different if the test is reliable － in theory.

Validity and reliability

Validity

According to Owen (1997), we may take into account two areas of inquiry while 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.

In defining validity one may consider what validity is intending to measure and to what degree it does so accurately. Cohen et al. inform us that "validity is an important key to effective research (and that) if a piece of research is invalid it is worthless" (2000:105). In addition, Cohen et al. continue that it is "impossible for research to be 100% valid" (2000:105) and regarding this knowledge we should consider 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). Considering Baker's (1989) model with regards to testing one could suggest that "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 that 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). 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 obtaining this accurately.

Several terms exist in categorizing the various ways of measuring validity. The following four categories exemplify the model illustrated by Hughes (1989) and Bachman (1990), these being construct validity, content validity, criterion-based validity, and face validity. Within content validity exists the variables of internal and external validity. These are also considered.

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).

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 exists 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. Cohen et al. elaborate stating that "internal validity seeks to demonstrate that the explanation of a particular event, issue, or set of data which a piece of research provides can actually be sustained by the data (and that) external validity refers to the degree to which the results can be generalized to the wider population, cases or situation" (2000: 107).

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 "demonstrated strong predictive validity" (2000: 112). 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).

Another important term related to validity is 'face validity'. This term relates to what degree a test is perceived to be doing what it is supposed to (Hughes 1989:27). In general, face validity in testing describes the look of the test as opposed to whether the test is proved to work or not.

Messick's 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 (cited in Hughes 1989). Messick considers these sole elements to be inadequate (cited in Bachman 1990) 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).

Considering the above framework defining validity in testing, one may consider the importance of determining what is appropriate for our own 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 can initially find focus in something that we can generally agree is an important outcome, the result.

Reliability

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, according to Hughes, that the higher amount of similarity there is between tests, the more reliable they would appear to be (Hughes:1989). Bachman (1990) adds that though 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 (Bachman 1990:164). Hughes is resolute in informing us that while we may attempt to make a test more reliable, we must try not to compromise its validity (2003:50). This is indeed a complex position. Balancing validity and reliability in various testing scenarios may be difficult as conditions are very rarely consistent. It may be important instead to identify what negative and positive consequences there are from testing rather than if there is a balance between reliability and validity.

Investigating reliability can also be approached by analyzing a test candidate’s Classical True Score (CTS). Hughes (2003:36) asks whether or not we can ever truly rely on test scores as a proof of reliability. 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. Certainly, it would seem, there is some division in the significance of scoring and its usefulness in calculating the reliability of tests. 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 (Bachman 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 there own ways of examining the source of error in testing.

Ensuring validity and reliability

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 out 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.

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 first hand experience of what may be required throughout the entire process of test organisation.

Method

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 forth week of each month respectively. The test appears in the appendices section.

The test is one of several listening tests conducted each term and is administered over the period of two weeks for approximately five hundred 1st grade students. Ten native speaking English teachers are involved in the design and administration, and marking of the tests. The eventual score is added to an overall score which is part of the students' final yearly grade, and is integral to individuals fulfilling requirements 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'.

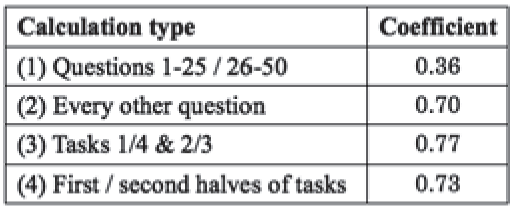
Split-half analysis

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.

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)) appear to suggest that there is a certain amount of unreliability between the two halves of the test. With the coefficient score of .36 measured one might suggest there are certainly opportunities to make salient problems within the test. Sections within the test were balanced so as to attempt to create relatively high equilibrium. 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, alternative ways of splitting the test was employed. The reliability coefficient was further analysed after collecting odd and even scores, from calculating various test task groups together, and a calculation drawn from 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



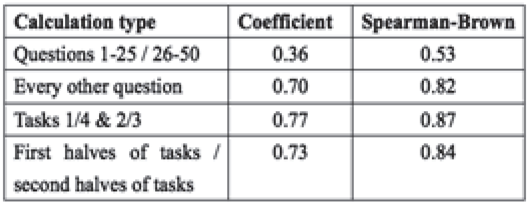
Calculation type	Coefficient
(1) Questions 1-25 / 26-50	0.36
(2) Every other question	0.70
(3) Tasks 1/4 & 2/3	0.77
(4) First / second halves of tasks	0.73

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:

Calc (4) ÷ Coefficient total (2.56) = <u>0.64</u>

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 x 2 = 1.28 / 0.64 + 1 = 1.64)

= 1.28 ÷ 1.64 = <u>0.78</u>

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

Qualitative results

In design, the four sections of the test were positioned to mirror each other to compensate for the subsequent split-half analysis (see Appendix). Each task was evaluated using the model described by Hughes (1989) concerning construct validity, criterion-related validity, content validity, and face validity. 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 endeavor to prepare the students for a post-course homestay in the UK which they are expected to utilise their language skills. 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 recognise 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.

With regards to 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.

Pedegogical consequences

Through paying attention to the various ways of creating positive tests, we can start to provide our students with testing that is suitable and appropriate for their progression as language learners. In conducting this research, it has been a valuable source of information in that it raises understanding of what is required to make tests efficient and consistent. Through knowledge of the techniques that assist in ensuring validity and reliability in the production, administration, and scoring of tests it has been an invaluable lesson of the complexities that exist and of how improvements can be applied to my own teaching situation.

Conclusion

Chapelle (1999) points out that “the challenge of moving from theoretical ideals concerning validation to specific practices that come into play in second language classes, programs, and research has been identified as ‘one of the critical challenges for testing professionals’” (1999:264). There may be some justification for small-scale analysis in that it can contribute in some way to general testing evaluative practices. Chapelle continues to say that validation of a test will not ultimately come from a single set of results but from “multiple sources of information” and that is what validation processes are intending to combine. The practical application of knowledge gained through analysis of our own teaching environments seems imperative and could contribute to improving testing on a wider scale. 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)

References

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

Bachman, L. and 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 Teachin*g (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.

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

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.

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

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

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

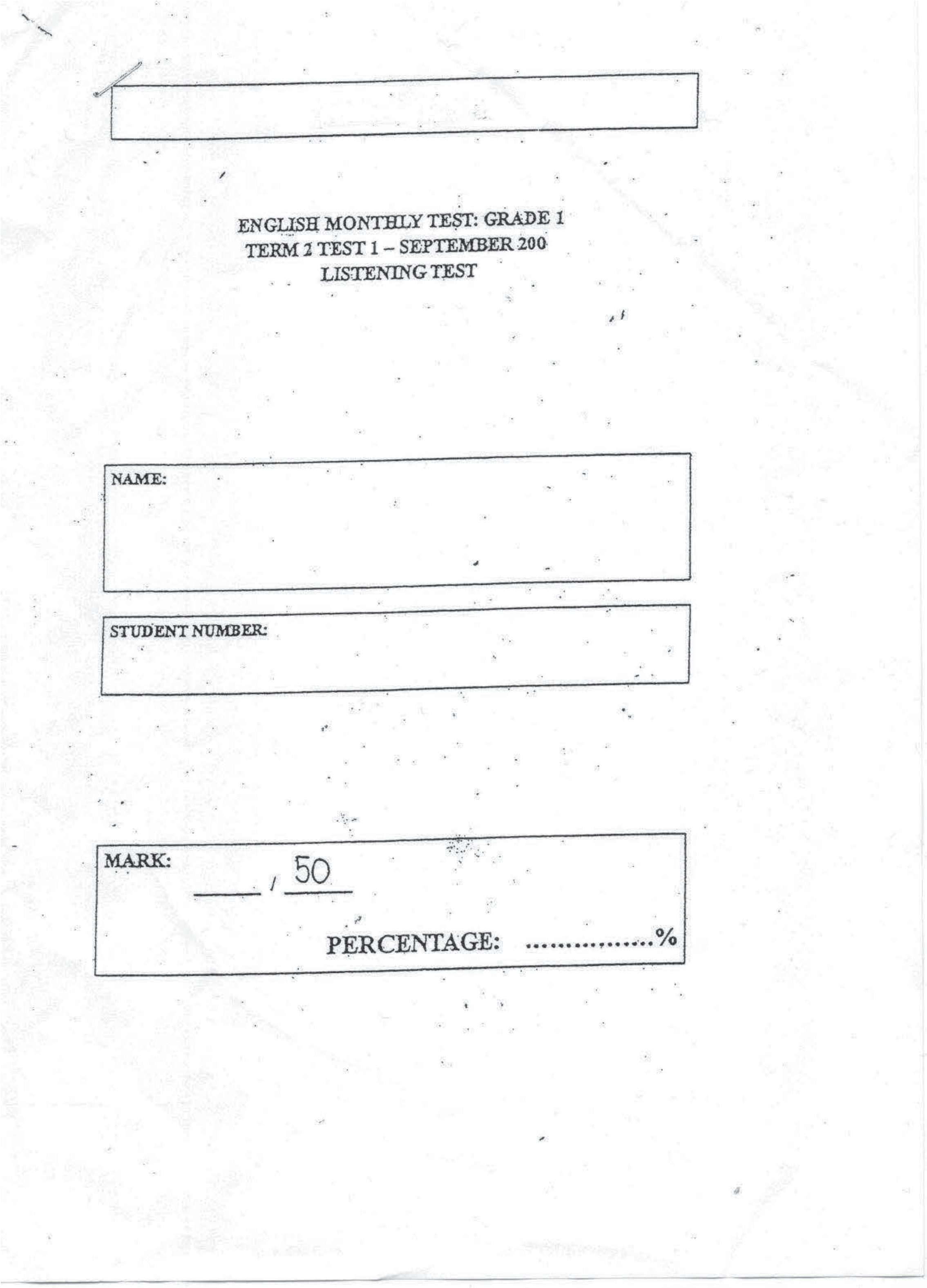
Student	Score 1 (25)	Score 2 (25)	50pts = 100%	Variability
1	24	25	49 (98%)	1 (1%)
2	20	22	42 (84%)	2 (4%)
3	21	23	44 (88%)	2 (4%)
4	25	25	50 (100%)	0 (0%)
5	24	25	49 (98%)	1 (2%)
6	21	23	44 (88%)	2 (4%)
7	24	21	45 (90%)	3 (6%)
8	25	24	49 (98%)	1 (2%)
9	22	23	45 (90%)	1 (2%)
10	21	21	42 (84%)	0 (0%)
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



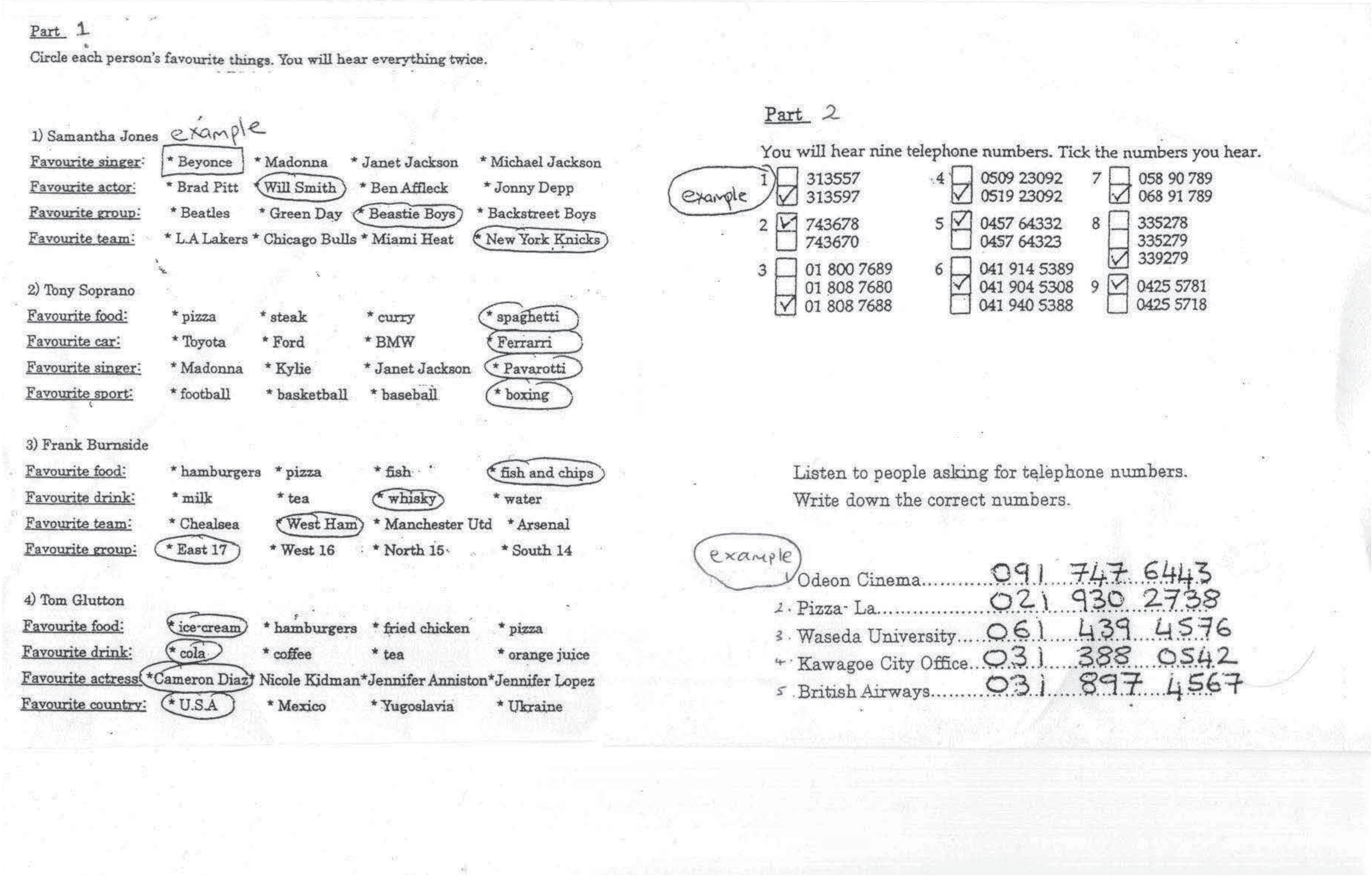
ENGLISH MONTHLY TEST: GRADE 1
TERM 2 TEST 1 – SEPTEMBER 200
LISTENING TEST

NAME:

STUDENT NUMBER:

MARK: ____ / 50 PERCENTAGE: ....%

Appendix 6



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:	* L.A 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

2 ☑ 743678
□ 743670

3 □ 01 800 7689
□ 01 808 7680
☑ 01 808 7688

4 □ 0509 23092
☑ 0519 23092

5 ☑ 0457 64332
□ 0457 64323

6 □ 041 914 5389
☑ 041 904 5308
□ 041 940 5388

7 □ 058 90 789
☑ 068 91 789

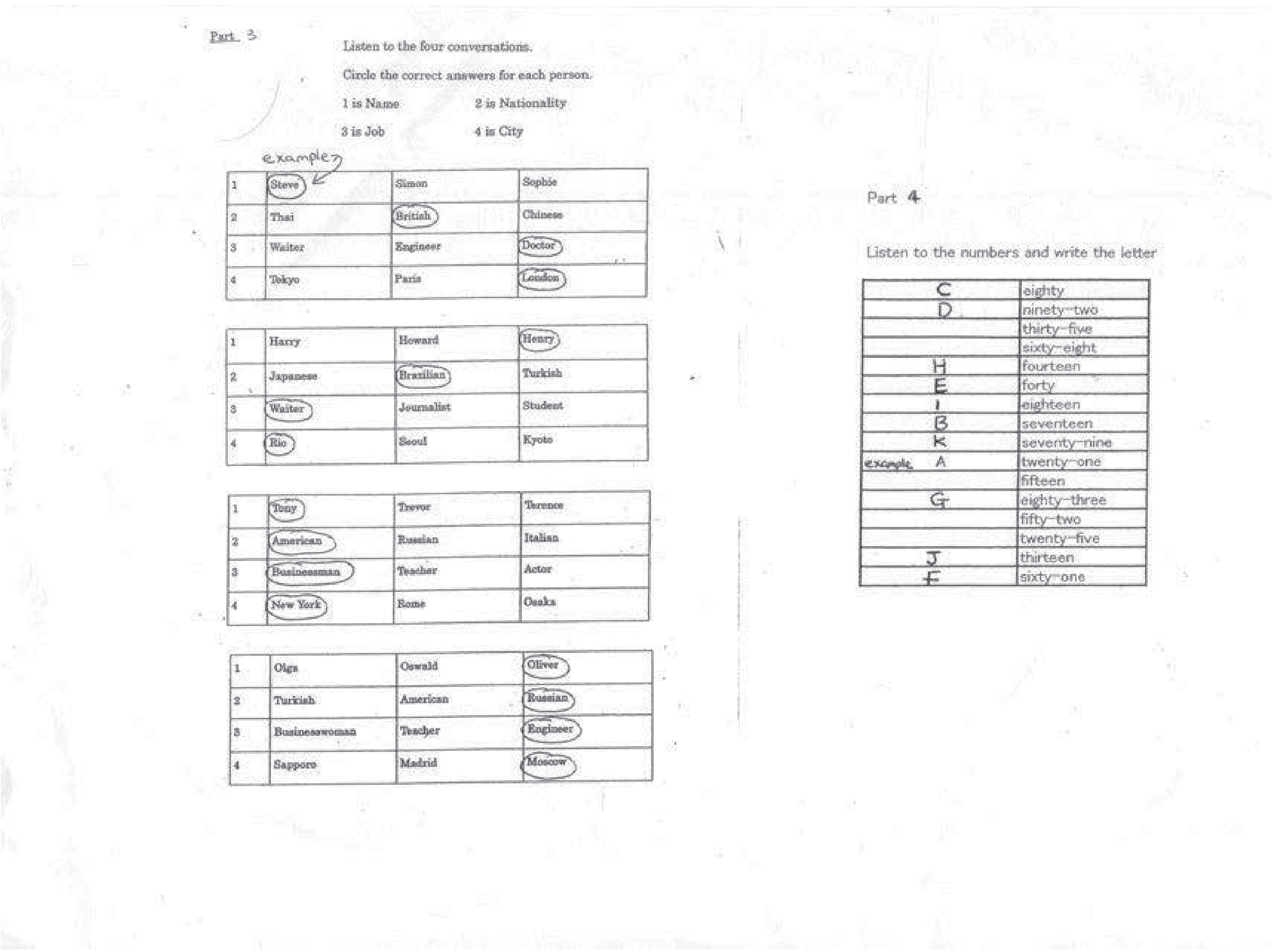
8 □ 335278
□ 335279
☑ 339279

9 ☑ 0425 5781
□ 0425 5718

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

example
1. Odeon Cinema.......... 091 747 6443
2. Pizza- La.......... 021 930 2738
3. Waseda University.......... 061 439 4576
4. Kawagoe City Office.......... 031 388 0542
5. British Airways.......... 031 897 4567

Appendix 7



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

1	[✓ Steve]	Simon	Sophie
2	Thai	[✓ British]	Chinese
3	Waiter	Engineer	[✓ Doctor]
4	Tokyo	Paris	[✓ London]

1	Harry	Howard	[✓ Henry]
2	Japanese	[✓ Brazilian]	Turkish
3	[✓ Waiter]	Journalist	Student
4	[✓ Rio]	Seoul	Kyoto

1	[✓ Tony]	Trevor	Terence
2	[✓ American]	Russian	Italian
3	[✓ Businessman]	Teacher	Actor
4	[✓ New York]	Rome	Osaka

1	Olga	Oswald	[✓ Oliver]
2	Turkish	American	[✓ Russian]
3	Businesswoman	Teacher	[✓ Engineer]
4	Sapporo	Madrid	[✓ Moscow]

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