Communicative Performance Opportunities & Proficiency

Colin Painter

This paper explores the relationship between communicative performance opportunities and proficiency as reflected in performance test scores. It also illustrates how the continuous assessment of oral communication performance in classes of Japanese university students was facilitated using multimedia computer software. First year learners worked in groups in a multimedia CALL laboratory. At their own pace and level, learners selected CD-Rom based video clips, predicted then practised communicative content, identified communicative aims, then employed them in self-created situations and requested assessment. With learners engaged in tasks, the teacher was able to supply pedagogic assistance and conduct testing. Criterion-referenced performance tests arise naturally from a functionally based course. In this study twenty-five tests were created and presented on role-cards. Learners indicated readiness for testing as they completed each unit of communicative activity. Three-minute tests focused on the communicative aim and thus the functions of the unit. Communication gaps made communication meaningful. A test role-card supplied the functionally based task, such as checking onto an airline flight. With the same group partners, learners acted out the task in pairs while the teacher listened and scored. With twenty-five units, and a test for each, learners and teacher were constantly aware of how well a language function had been assimilated and how well learners were getting used to communicating in meaningful situations. The validity of such tests is good and judging by the results of repeat tests, when they occurred, so also is reliability. Moreover, there appeared to be a significant relationship between performance opportunities and proficiency.

Background

A previous study into the development of oral communication using computers (Painter, 1995) showed how paired learners requested testing through role play after they had completed a unit of functionally-based language activity. The use of role-play where sets of instructions are given to partners in a situation which requires an exchange of information has been covered in the literature by Underhill (1987), Hughes (1989), Seliger and Shohamy (1989). Typically, an information gap requires the participants to accomplish a task by exchanging information. The ability to do this is then a measure of communicative success. However, researchers suggest test designers should be careful to avoid making the test solely a problem-solving task. A person who can think logically and quickly might do well on the test but the result may not reflect language ability. Furthermore, role-plays can be a welcome support to learners who are afraid of having nothing to say while being a constriction for those who wish to express their own opinions. Hughes (1989) feels that where learners are to interact with peers, the performance of one learner is likely to affect that of the other. He proposes matching candidates if possible. Concerning the method of eliciting the performance, Underhill suggests that pairing learners dispenses with inhibiting factors which may be present in a pairing of interviewer and learner. Procedures should be explained clearly beforehand and the role-play described in writing. However, understanding the instructions should not become a part of the test. Among other types, Underhill suggests functions as a subject of such role-play situations. Well documented functional outline sources are Wilkins (1973, 1976) and Van Ek (1975). However, as Hughes points out, the functions targeted in a specific test might not get elicited during the test. Additionally, the ability of the learner to adopt an identity is necessary for success and sensitivity is needed concerning the likely presence of any cultural

barriers to such an adoption. Selecting names in the role-play which are not gender-specific could reduce problems when roles are assigned. Underhill suggests that learner-learner role-plays produce more involvement and greater spontaneity than teacher-learner types however there is a need for time limits. With regard to marking, Underhill presents two mark categories, the traditional and more modern. In the traditional appear: Grammar; Vocabulary; Pronunciation/ Intonation/Stress; Style/Fluency, Content. In the more modern "Performance Criteria" (derived from Carroll, 1977) appear: Size (length of utterance); Complexity (attempts at complex language); Speed (speaking speed); Flexibility (ability to adapt to changes); Accuracy; Appropriacy; Independence; Repetition; Hesitation (p96). Underhill stresses that one assessor can only track three or four of these categories simultaneously. Any required weighting of marks could be done at a later stage. Hughes mentions the analytic "criterial levels of performance" categories (P102) of the RSA (Royal Society of Arts) test, for oral interaction at the intermediate level, which reflect to some extent the "Performance Criteria" above although he feels that they are not very precise. He also illustrates a holistic approach where content is combined with "criterial levels of ability" (p103) in examples from the ACTFL (American Council for the Teaching of Foreign Languages) Guidelines and the IRL (Interagency Language Roundtable) ratings for similar levels. Underhill's comments on subtractive marking are worth noting. He feels that such marking causes the assessor to focus on mistakes, sufficient for accuracy alone but not leading to a fair judgement of spoken proficiency.

Intra-marker reliability is important and inter-marker reliability may also become an issue if more than one marker is involved. Test reliability will not be so high where more subjective judgements are required. Indeed, some testers avoid subjective tests. However, while objective tests are easy to mark, it is doubtful that they are valid enough for testing an

integrated skill such as communicative performance. Some tests attempt to combine the subjective and objective element thus combining an amount of validity and reliability. Hughes (1989) cites the FSI (Foreign Service Institute) oral test as an example of how analytic (objective) and holistic (subjective) approaches are combined. Research revealing the connection between objective and subjective scoring resulted in tables which are used to convert an analytic score to a holistic score. Hughes attests to their efficacy. Underhill (1987) feels that designing a subjective test of oral proficiency with claims to reliability is indispensable. Having more than one assessor would reduce the problem of reliability. Finocchiarro and Brumfit (1983) view test reliability as a lot less important than test validity. Similarly, Underhill feels that classical measures of test reliability have little relevance for oral tests since they were designed for the more rigid tests which yield either correct or incorrect answers. Assessors will gain more useful information by designing their own systems for comparing scores across markers. In the context of oral testing, Underhill prefers to see reliability as a specific form of general validity. Although he notes that reliability is usually perceived as a different concept from validity. He maintains that a test cannot be generally valid unless it is reliable. The important question to ask is whether the test does what it is supposed to do. The question concerning reliability can supply one kind of answer as do other specific forms of validity. Lado (1961) suggested that a good oral production test would usually have a reliability coefficient range of .70 - .79. For comparison, tests of reading and comprehension would be .90 - .99 and .80 - .89, reflecting the relative difficulty of attaining reliability in oral production tests. Davies (1978) referred to the validity-reliability "tension" where striving for test reliability can reduce the validity and vice-versa. Davies (1968) presented five kinds of validity: Face, Content, Predictive, Concurrent and Construct. However, Bachman (1990) suggests that face validity is no longer a standard for validity.

Test validation is related to how tests will be used and evidence is collected to support the way in which a test is used. According to Bachman this evidence can be grouped in three categories; content relevance, criterion relatedness and meaningfulness of construct. Although these have been treated elsewhere as different types of validity, usually named content, criterion and construct, they are complementary. Brown (1988) concurs with these categories. Morrow (1979) points out that excepting face and perhaps predictive validity the others are circular. Assumptions about the nature of language and language learning will produce tests which are valid in terms of the assumptions but the tests are devalued as soon as the assumptions are questioned. The characteristics which Morrow expects for a test of communicative ability (p150) are reproduced below:

1. It will be criterion-referenced against the operational performance of a set of authentic language tasks. In other words it will set out to show whether or not (or how well) the candidate can perform a set of specified activities.
2. It will be crucially concerned to establish its own validity as a measure of those operations it claims to measure. Thus content, construct and predictive validity will be important, but concurrent validity with existing tests will not be necessarily significant.
3. It will rely on modes of assessment which are not directly quantitive, but which are instead qualitative. It may be possible or necessary to convert these into numerical scores, but the process is an indirect one and recognized as such.
4. Reliability, while clearly important, will be subordinate to face validity. Spurious objectivity will no longer be a prime consideration, although it is recognized that in certain situations test formats which can be assessed mechanically will be advantageous. The limitations of such formats will be clearly spelt out, however.

Performance is an integrated occurrence and testing isolated discrete items will demolish this integrity. Morrow feels that for this reason quantitative methods are impractical and qualitative methods should be used. He notes that this in turn affects reliability. He argues that to test proficiency, regardless of how refined the parts may be, it is impossible to obtain an actual measure of language performance from tests of the parts alone. He concludes that performance tests have most value in a communicative context. Morrow also concurs with the mark categories termed "Performance Criteria" above, as derived from Carroll (1977). He envisages that the pass/fail concept will have less value and performers will be assessed in terms of what they can do. For administrative purposes a particular level may be required for a grade mark but even low scorers can be told what they have achieved.

Another important aspect of testing is the backwash effect. To have a beneficial effect on learning and teaching, an oral performance test should encourage the ability being tested, that is: oral performance. It would also test directly and specifications should be criterion-referenced. Achievement and progress tests should be based on objectives, rather than teaching and the content of textbooks. Where the syllabus and teaching match the objectives, tests based on these objectives will more accurately measure learning and teaching, producing a beneficial backwash. It follows that test forms such as cloze will have a less beneficial washback effect, however, they can provide correlational assistance. Researchers, such as Oller (1973), indicate that cloze tests can indicate a basic level of language proficiency, although obviously they are unable to directly indicate a testee's ability to perform in a language. In one study, Geva (1992) reported a high correlation ($r = 0.69$, $p = <.001$) between oral proficiency ratings (using an FSI type instrument) and the cloze and suggested that both cloze and oral proficiency ratings may tap a general L2 discourse proficiency

factor.

Reliability and validity can be analysed through statistical studies. However, as Brown (1988) points out, of the two main categories of language test; norm-referenced and criterion-referenced, the latter is less accommodating to statistical study. A criterion-referenced test, as shown above, is typically used to measure what learners have achieved with reference to a criterion level which defines the ability objectives of a unit of study or of a course of study. It is therefore conceivable that if learners have successfully learnt the ability they could all score full marks. However, without a dispersion of scores, statistical methods have little use. As Bachman (1990) points out, reliability estimates depend on the amount of variability in test scores. He hypothesizes a situation where an achievement test is administered to learners based on course objectives, at the beginning of a course, again after two weeks and again at the end. At the beginning, the scores would likely be uniformly low. After two weeks, if instruction had been effective, scores would be slightly higher but with little variation between them. Again, at the end, assuming instruction to be equally effective for all, we could expect uniformly high scores, and again, little variation between them. A statistical estimate of internal consistency would probably yield low reliability coefficients. Likewise correlation of the first two sets of scores would probably yield a very low estimate of stability due to the little variance in test and retest scores. However, as Bachman suggests, by intuition it is obvious that scores may accurately reflect changes in the achievement of content objectives. For this reason classical norm-referenced estimates of reliability are ineffective with criterion-referenced test scores. In answer to this, Bachman refers to the defined set of tasks, from which a test is drawn, as a "domain". In criterion-referenced tests the interpretation of the "domain score" represents a learner's level of achievement in terms of the domain of ability criteria. Whereas in norm-referenced tests the

learner's score would be interpreted in terms of the average performance of a group of learners. The "domain score" is to criterion-referenced tests what "true score" is to norm-referenced tests. The term "reliability coefficient" has been rejected by some researchers in the context of criterion-referenced tests in favour of what Berk (1984) calls an "agreement index" and Kane and Brennan (1980) call a "dependability coefficient". In criterion-referenced tests, reliability is concerned with: (i) The dependability of test scores as indicators of learners' ability level in a given domain, (ii) The dependability of decisions made on the basis of criterion-referenced test scores. According to Bachman (1990), the method of calculating domain score dependability involves use of the Kuder-Richardson (1937) formula 20 (KR-20) which involves computing the means and variances of dichotomous test items. This is then used to calculate the reliability coefficient. Brown (1989) has evolved a more practical formula for dependability requiring less computation than Kane and Brennan's dependability coefficient. Brown's method of calculating domain score dependability also involves use of the KR-20.

The Current Study Purpose

The purpose in this study was to explore the relationship between communicative performance opportunities exploited by learners and proficiency as reflected in performance tests. It was also appropriate to establish reliability and validity for the testing. In the process it was hoped to illustrate further how the continuous assessment of oral communication performance was facilitated using multimedia computer software (Milward, 1993). Information from learner evaluation of the program was included in the interpretation of results.

Test Development

A test would deal with one segment of study containing a defined communicative aim and functional definitions. Twenty-five pairs of role-cards were produced which outlined a situation and a task. A communication gap was embedded in the design of the task. As learners completed each unit of communicative activity, they identified the communicative aim and indicated readiness for testing. The tests, approximately three-minutes in duration, focused on the communicative aim and thus the functions of the unit. Successfully accomplishing the test task would signify achievement of the communicative aim and of a performance criteria. Each learner, in a pair of testees, was issued with one of two role-cards. Testee pairs were synonymous with learner pairs and the task was acted out while the teacher listened and scored. There was, however, the possibility of one learner negatively affecting the performance of the other (Hughes 1989). Error by one testee could confuse the other testee. In this case the assessor would supply the correct exponent. This action, in turn, could have an effect on the testees. For this reason, the assessor would supply the information in a non-judgmental manner taking pains to discount any implication of seriousness in the intervention. In practice it would take the form of a spontaneous remark simply designed to keep the conversation flowing. Procedures were explained in writing and verbally at the beginning of the course. Practically speaking, learners could participate in many tests, therefore the majority of tests were taken by seasoned test-takers. In this way any adverse effect of the form of the test diminished although other aspects, such as predicting content, could augment. A functional basis for the test would imply a similar origin for the syllabus of the course, thus fulfilling Davies' test content validity (Davies 1968); that a test should accurately reflect the underlying syllabus. Wilkins (1972, 1974) and Van Ek (1975), among others, provide a possible

framework for such a design in their formulation of the Threshold Level which was intended as an international standard level for language learning. The Threshold Level might be generally compared with a lower intermediate level. Wilkins and Van Ek's notional-functional categories evolved from consideration of the situations in which learners would have to use a foreign language, the roles the learner would have to play, the settings and the topics which would have to be handled. In Table 1.1 and 1.2 the functional categories exploited in the two levels of the present study are displayed against the title of each topical unit. The six main categories of verbal communication intended for the Threshold Level (Van Ek, 1975) are illustrated in Table 2. For comparison, alongside the categories, are the numbers of the units which contain corresponding functions covered by the learners in the present study. Units have been matched against the functions of primary use. However, when considering secondary use, the units fit equally with multiple function categories.

Table 1

Outline of Course Functions

Level One

Unit	Title	Functions
1-01	Introduction:	introduce self & discuss itinerary/ purpose, describe possessions
1-02	Information	express/inquire about wants/ preference, inquire about availability & request further information
1-03	Food	express/inquire about wants/ preference, inquire about availability & request further information & choose
1-04	Home	identify relationship/ownership, express pleasure/liking
1-05	Inclusive	ask about/describe occupation & offer/request refreshment

Table 1. 2

Outline of Course Functions

Level Two

Unit	Title	Functions
2-01	Arrival	asking／giving personal information
2-02	Information	finding satisfactory accommodation
2-03	Hotel	checking-in／giving information
2-04	Restaurant	complaining
2-05	Bar	discuss intentions／plans
2-06	Estate Agency	describing location
2-07	Apartment	talk about lifestyle／accommodation
2-08	Appliance Shop	discuss habits／routines
2-09	Home	talk about a sequence of past events
2-10	Telephoning	discuss who you know／remember／forget
2-11	Telephoning	discuss quantity, duration, distance
2-12	Post Office	ask／explain procedures
2-13	Restaurant	compare／evaluate things done／seen
2-14	Clothing Shop	talk about wants concerning undetermined object／quantity／person／place
2-15	Pharmacy	explain／advise someone with a problem
2-16	Home	talk of things done／seen
2-17	Bookshop	compare things／people／places
2-18	Cafe	talk about intentions／want／desire, periods of time past／future
2-19	Bank	talk about getting things done／things already done／accomplished
2-20	School	interviewing／talking about past／what was happening at a given time

Table 2

Comparison of Threshold Functions & Course Functions

	Threshold Function	Course Function	
		Level 1	Level 2
1	Imparting and seeking factual information	1, 4	1, 2, 3, 6, 7, 8, 9, 16, 20
2	Expressing and finding out intellectual attitudes	3, 5	10, 19
3	Expressing and finding out emotional attitudes	2, 3, 4	2, 4, 5, 13, 14, 17, 18
4	Expressing and finding out moral attitudes		4, 14, 17
5	Getting things done		1, 12, 15, 19
6	Socialising	1	

The tests in the present study took the form of situations in which learners played roles in particular settings concerning particular topics. For example the situation in Level 2, Test 1 (Appendix A), put testees in the roles of: receptionist and patient, within the setting of: a hospital, and a topic of: seeking medical attention. The task required the

receptionist to request personal information for the records and the patient to explain the condition and request help. To successfully carry out the task the testees needed to perform the functions which had been practised, identified, and exponentially recreated at the study stage during lesson time.

The scoring principle had been indicated to learners in a procedure guide as follows:

1 communication was meaningful & grammatically correct:	2 points for each section
2 communication was meaningful but contained grammatical errors:	1 point for each section
3 communication was meaningless:	0 point for each section

The scoring method attempted to reduce the number of items the assessor needed to keep track of during the test (Underhill, 1987). The method also attempted to reduce the need and influence of subjective judgement and help keep the functional target in focus. Results were announced to individual testees at the end of the test.

Method

Subjects

24 mixed gender 1st year learners in a class enrolled alphabetically in the Faculty of Administration. The time period was one academic year and the English language CAI class frequency was once a week; totalling 26 classes (13 in each of two semesters). Classes were 1.5 hours each with a total of 39 hours for the year.

Instrument & Procedure

Reliability

The present performance tests did not employ dichotomous test items, that is, items which could be scored right or wrong. Therefore dependability could not be calculated using the KR 20 formula. Similarly, item analysis could not be employed since items were interdependent, each pair of testees generating unique question and answer content.

Test-retest data, shown in Table 3.1, was examined for normal distribution, equal variance and linearity. Stability (test-retest reliability) was estimated using 9 pairs of tests scores from repeated tests and calculating the reliability coefficient with the Pearson product-moment correlation coefficient. Intra-rater reliability was also indicated by the same correlation. Results appear in Table 3.2. During a test the assessor would not be aware of the test status, i.e., first test or retest.

Table 3. 1

Performance Test-Retest Data

Level	Test %	Retest %	Interval in weeks
1-1	80	80	2
1-2	60	70	1
1-3	80	70	15*
2-1	100	100	2
2-2	60	80	15*
2-2	80	90	15*
2-3	80	90	11*
2-3	80	90	11*
2-3	100	100	11*
m	80	90	
SD	13.3	10.7	

* = includes 10 week summer break

Table 3. 2

Performance Test-Retest Correlation

Dependent (X) & Independent variable (Y)	r	r^2
(X) Performance Test Scores & (Y) Retest scores	0.88	0.77

$p < .05$, $df = 7$.

Validation

1. Concerning content validity, (i) the ability domain was based on the functional course outline; (ii) test method facets (the setting and procedure) were evaluated and (iii) the degree to which test task represented the ability domain was evaluated. This evaluation was facilitated by the specific focus and limited nature of tests.

2. Criterion validity, implies correlation with a validated test and is here subsumed under construct validity.

3. Construct validity, is operationalized here with 'construct' as: the proficiency to perform in a defined language function area. Learners in the current study were additionally given two cloze tests, one in each semester. Performance test score and cloze score scattergrams were examined for normal distributions and linearity, The Pearson product-moment correlation coefficient was used and the results are shown in Table 4.

Performance Quantity & Performance Score Correlation

The two interval scales of performance scores and performance quantity (how many tests learners sought to take) were analysed for correlation. The Pearson product-moment correlation coefficient was used and the results are shown in Table 5.

Evaluation

An evaluation was conducted at the end of each semester, using the same instrument in both cases, the full instrument has been shown in a previous paper (Painter, 1995). Relevant information concerning testing is presented in the results. 24 learners supplied information anonymously.

Table 4

Performance Score & Cloze Score Correlation

Dependant (X) & Independent variable (Y)	r	r^2
(X) Performance test scores and (Y) Cloze:	0.62	0.39

$p < .05$, $df = 22$.

Table 5

Performance Quantity, Performance Score & Cloze Score Correlation

Dependant (X) & Independent variable (Y)	r	r^2
(X) Performance quantity and (Y) performance scores:	0.41	0.17
(X) Performance quantity and (Y) Cloze:	0.51	0.26

$p < .05$, $df = 22$.

Results

Test-retest Reliability

In the performance scores test-retest correlation study (Table 3. 1 & 3. 2), the correlation coefficient $r = 0.88$, was significant at $p < .05$, $df = 7$.

The coefficient of determination r squared = 0.77. The estimate for intra-rater reliability results from the same correlation coefficient, $r = 0.88$ significant at $p < .05$, $df = 7$. This correlation gives a positive and significant estimate of test stability. Likewise intra-rater reliability is high.

Validity

In the performance test score and cloze score correlation study (Table 4) the coefficient indicates a correlation of $r = 0.62$, significant at $p < .05$, $df = 22$. This estimates a medium correlation between the two variables concerning construct validity.

Performance Quantity & Performance Score Correlation

The correlation coefficient of $r = 0.41$ between performance quantity and performance scores is low and indicates a weak but significant relationship between the two variables at $p < .05$, $df = 22$. The correlation coefficient of $r = 0.51$ between performance quantity and cloze scores is fairly low and indicates a weak to medium, significant relationship between the two variables at $p < .05$, $df = 22$. The coefficient of determination, r squared, estimates the extent to which the two variables overlap. 17% of the variation in performance scores is due to the variation in performance quantity. 26% of the variation in cloze is due to the variation in performance quantity.

Evaluation

Learners were asked specifically whether measuring their oral English ability in the computer laboratory was effective or not. Learners answered on a scale of 1-5, low-high estimate. The resulting 'means' were,
1st semester = 3. 58 and 2nd semester = 3.79.

Conclusion

Results of test-retest reliability and intra-rater reliability present high estimates of stability and suggest tests were reliable. The correlation of performance quantity and cloze score was also significant and offers a fair estimate of construct validity. Along with content validity this suggests reasonable confidence in test validity.

The weak but significant estimate of the relationship between performance quantity and performance scores is interesting. That 17% of the variation in performance scores is due to the variation in performance quantity may be grounds for further investigation. The estimated closer relationship, of 26%, between performance quantity and cloze also sustains the idea that performance quantity does support underlying aspects of proficiency.

In the present study an attempt was made to provide learner autonomy in the belief that self direction would encourage production as well as provide the opportunity to learn at the appropriate level. Within a single ninety minute period, once a week, it may be difficult for the average learner, in a group of twenty-four, to maintain serious interest. The limited time available to learners to exercise their autonomy to perform was perceived as a constricting factor. In spite of this, learners' perceptions of the effectiveness of measuring their ability increased.

From the perspective of testing, with an average of eight tests taking place per lesson in addition to pedagogic assistance, learners sometimes had to compete for the chance to test, possibly dampening the positive effects of autonomy. On the other hand, learners benefitted from immediate knowledge of their assessment rather than having to wait until the end of the semester. Assimilation of learning and whether the tests, following soon after practice, could measure assimilated ability, needs further investigation. However, the washback effect of such testing was positive.

To learn the real significance of the relationship between performance opportunities and proficiency it would be necessary to provide truly unconstricted opportunity. Without this, the possibility of learning the real significance of the relationship may be reduced. Further research could include self testing by learners. In this way, learners would be able to progress without any impediment caused by the test event.

In spite of possible constrictions on learner performance in this study the evidence suggests that proficiency is significantly related to performance opportunities.

Acknowledgements

This is a version of a paper presentation at the Japan Association of College English Teachers (JACET), 35th Annnal Convention Program, Kyoto, Japan. I would like to thank Dr. John Shillaw of the Language Centre, Tsukuba University, for his valuable reading and comments. I am also grateful for comments from Dr. Thomas Robb, Chairman, English Dept., Kyoto Sangyo University. In addition I would like to thank all

the students who participated and provided data, without whom the study would havebeen impossible. I am also grateful to colleagues for their support. Errors remain my own.

The Author

Colin Painter is Associate Professor at the Prefectural University of Kumamoto. He has taught at universities in Asia for the last 14 years. His interests include research in language acquisition and teaching, curriculum development, and computer assisted language learning.

References

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

Berk, R. A. (1984). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp 231-266). Baltimore, Md.: The John Hopkins University Press.

Brown, J. D. (1988). *Understanding research in second language learning,* Cambridge: Cambridge University Press.

Brown, J. D. (1989). *Short-cut estimates of criterion-referenced test reliability.* Paper presented at the 11th Annual Language Testing Research Colloquium, San Antonio, March 1989.

Carroll, B. J. (1977). *Specifications for a new English language exam.* London: Royal Society of Arts, mimeo.

Davies, A. (1968). (ed.) *Language testing symposium,* Oxford: Oxford University Press.

Davies, A. (1978). Language testing. *Language Teaching and Linguistics Abstracts, 11,* 3-4.

Finocchiarro, M. and Brumfit, C. (1983). *The functional-notional approach.* Oxford: Oxford University Press.

Geva, E. (1992). The role of conjunctions in L2 Text Comprehension, in *Tesol Quarterly, 26* (4), 731-747.

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

Kane, M. T. and Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychology Measurement 6,* 125-160.

Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika 2,* 151-60.

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

Milward, M. (1993). Nova City. (CD-ROM software) Tokyo: Nova Information Systems.

Morrow, K. (1979). Communicative language testing: revolution or evolution? In C. J. Brumfit and K. Johnson (Eds.), *The communicative approach to language teaching* (pp 143-157). Oxford: Oxford University Press.

Oller, J. (1973). Cloze tests of second language proficiency and what they measure, *Language Learning, 23* (1).

Painter, C. (1995). Developing oral communication using computers: computer assisted language learning. *Administration, 2*(3), 109-150.

Seliger, H. W. and Shohamy, E. (1989). *Second language research methods.* Oxford: Oxford University Press.

Underhill, N. (1987). *Testing spoken language.* Cambridge: Cambridge University Press.

Van Ek, J. A. (1975). *The Threshold level.* Oxford: Pergamon Press.

Wilkins, D. A. (1972). Grammatical, situational and notional syllabuses, *Proceedings of the Third International Congress Of Applied Linguistics, Copenhagen 1972,* Heidelberg : Julius Groos Verlag,

Wilkins, D. A. (1973). *An investigation into linguistic and situational content of the common core in a unit-credit system.* Strasbourg: Council of Europe.

Wilkins, D. A. (1974). Notional syllabuses and the concept of a minimum adequate grammar, in Corder and Roulet (Eds.), *Linguistic insights in applied linguistics,* AIMAV／Didier.

Wilkins, D. A. (1976). *Notional syllabuses.* Oxford: Oxford University Press.

Appendices

Appendix A

Level 2 Test 1

Student A:

You are Jess Brown, a photographer living in New York.
You ate some food in a cheap restaurant last night but now you feel sick.
You have just arrived at the reception of Central Hospital.
You would like some medicine.

L2 01

Student B:

You are Jo Francis, a receptionist at Central Hospital.
When new patients arrive you must get their name, address,
profession and age.
You should then tell them to sit down and wait for the doctor.

L2 01