Lexical Density In Candidate Output On Two Versions Of An Oral Proficiency Test.

Kieran O'Loughlin

1. Introduction

This article compares candidate output in direct (live) and semi-direct (tape-mediated) versions of an oral proficiency test. This is one component of a four-skill English proficiency test currently administered to certain categories of intending immigrants to Australia. The test is designed to be taken by these people in their country of origin; two versions of the Oral Interaction sub-test are necessary for economy and flexibility, since the human and/or technical resources available in each overseas test centre may mean that only one version can be administered. The test is known by the acronym access: or the Australian Assessment of Communicative English skills[1]. Specifically, the study contrasts candidate output on selected matching tasks in the two versions of the test from the perspective of *lexical density* which provides a measure of the relationship between lexical and grammatical items in spoken and written discourse. This is used as an index of the degree of 'orality' versus 'literacy' contained in the language samples collected. The findings have implications for the content validity as well as the interchangeability of the two kinds of tests.

2. Background to the study

Previous research on the comparability of direct and semi-direct tests of oral proficiency has focused mainly on the Oral Proficiency Interview (OPI), a direct test and its more recently developed semi-direct surrogate, the Simulated Oral Proficiency Interview test

[1] This test has been developed under the aegis of the National Centre for English Language Testing and Research (NCELTR) at Macquarie University, Sydney as a project sponsored by the Australian Commonwealth Department of Immigration and Ethnic Affairs (DIEA). The current study was also funded by DIEA. The author wishes to gratefully acknowledge the academic support of Professor Chris Candlin at NCELTR as well as Associate Professor Tim McNamara and Dr Gillian Wigglesworth at the NLLIA Language Testing Research Centre, University of Melbourne.

Melbourne Papers in Language Testing 1994 Volume 3.1 pp. 26-49 The Language Testing Research Centre, The University of Melbourne.

(SOPI), which are both widely used in a number of countries to assess oral proficiency in a variety of languages.

The OPI consists of a face-to-face interview by a trained interlocutor (who usually also carries out the assessment) and can include a role play segment. In general, the interviewer is free to ask whatever questions s/he wishes and the questions are different for each candidate. The topics and language input are adjusted according to the candidate's perceived proficiency. The SOPI, on the other hand, is a tape-recorded test which is invariant. Initially, as in the OPI, there is a 'warm up' phase where the candidate is asked a number of simple personal background questions. The rest of the test consists of a series of tasks which elicit oral discourse through the use of both aural and visual stimuli and the candidate's responses are recorded. Parts 2, 3 and 4 are pitched at the Intermediate and Advanced levels of the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines or Levels 1 and 2 of the Federal Interagency Language Roundtable (FILR) skill level descriptions while Parts 5 and 6 assess the candidate's ability to operate at the Advanced and Superior levels (or from Levels 2 to 4 of the ILR skill level descriptions). Both tests often end with a 'wind down' phase where candidates are asked one or two easy questions designed to relax them and to end the test as naturally as possible. Unlike the OPI, the SOPI is assessed retrospectively by trained raters using the audio-taped recording of the candidate's test performance. Both tests, however, are assessed holistically using the ACTFL/ILR scale. (Stansfield, 1991; Stansfield and Kenyon, 1992).

On the basis of research carried out at the Centre for Applied Linguistics (CAL) in Washington DC, Stansfield (1991) suggests that the SOPI has shown itself to be a valid and reliable substitute for the OPI. In relation to the issue of concurrent validity, he reports Pearson correlations of between 0.89 and 0.95 for the scores of candidates on the two kinds of tests in various languages. In a later study, Stansfield and Kenyon (1992) use generalizability theory to lend further support to this claim. Generally low levels of subject by test interaction were found for candidates who had undertaken the two kinds of tests, again in a range of languages.

In relation to the issue of validity, Stansfield (1991) suggests that one important problem with the OPI is that the candidate's

performance is in large part determined by the skill of the interviewer whereas the SOPI offers the same quality of language input to each candidate. This, he asserts, is a major consideration in choosing which type of format to use depending on the purpose of the test. His conclusion is that the OPI may be more suitable for placement and program evaluation purposes and the SOPI more appropriate when important decisions are to be based on test scores given the high degree of quality control it offers.

In addition, Stansfield (1991) argues that the reason why OPI and SOPI correlate so highly may be because neither, in fact, allows candidates to demonstrate their interactive skills. Even in the OPI, he contends, both interviewer and candidate understand that it is the examinee's responsibility to perform — little true interaction takes place (Stansfield 1991:205). However, this argument appears to be somewhat dubious since interaction is integral to any live test of oral proficiency. It is the nature of the interaction which needs to be carefully understood in determining what kind of language is being measured. The OPI is usually a dynamic speech event in which the interviewer makes a substantial verbal contribution throughout and both their language and the topics are aligned according to the level of proficiency the candidate appears to be operating on. Even if it is correct that an interview is very different from a "natural" conversation, the OPI is still a much more interactive language event than the SOPI in which the input from the "interviewer" is invariant.

On the basis of their research focusing on test scores, Stansfield and Kenyon (1992) conclude that the OPI and SOPI are highly comparable as measures of oral language proficiency: they may be viewed, it is asserted, as parallel tests delivered in two different formats.

Shohamy (1992) adopts a more sceptical position in relation to statistical comparisons of scores on the two kinds of tests. Unlike Stansfield (1991), she argues strongly that high correlations between scores on different kinds of tests provide necessary but insufficient evidence for test substitution i.e. they may not necessarily be measuring the same kind of language. Consistent with a post-positivist outlook, she underlines the need to examine the validity of tests from multiple perspectives, not just from the point

of view of test scores, to obtain a deeper understanding of what they actually measure.

In this study Shohamy (1992) reports findings from a discourse analytic study of the OPI and SOPI. She found that the SOPI elicited a more limited range of language functions than the direct version and that SOPI answers included more self-correction, repetition of phrases in the eliciting questions and paraphrasing. There was also a more restricted range of prosodic features in the SOPI, mainly hesitations and silence when no answer was available. The discourse produced in the SOPI was also more formal and cohesive. Furthermore, and of crucial relevance to this study, she found in the OPI that the relationship between the amount of lexicon and grammar was approximately 40% lexicon and 60% grammar, while these figures were reversed for the SOPI i.e. 60% lexicon and 40% grammar. This relationship is known as a measure of *lexical density* (usually expressed simply as the percentage of lexicon) and used as an index of the degree of 'orality' versus 'literacy' in both spoken and written discourse. Texts which are more literate — and these include both written texts and spoken texts such as speeches — will be characterised by a higher degree of lexical density (i.e.contain a higher percentage of lexical items) than more oral texts which include both spoken texts and written texts such as highly informal letters (Ure, 1971; Halliday, 1985).

Shohamy's interpretation of the findings from this study is that:

> *... the context of a test, 'face-to-face' versus 'tape-mediated' affects, or even dictates, the type of language that is produced. The physical presence of a human interlocutor on the OPI is probably what causes the production of language that is more conversational and more intimate, while talking to a tape on the SOPI produces 'tape-like' discourse which consists of a limited number of speech functions.*
>
> Shohamy (1992:20)

Shohamy (1992) concludes that while different discourse samples are obtained in the two tests, it is difficult to determine which language sample, or test, is better as new developments in communication technology (eg answering machines, dictaphones, e-mail) challenge the primacy of face-to-face talk. It may be, she

suggests, that a valid assessment of oral language proficiency ideally requires the use of both kinds of tests. Where a choice between the two tests formats needs to be made, a variety of factors should be considered including the context and purpose of the test.

Shohamy's (1992) study suggests that caution should be exercised in drawing conclusions about the interchangeability of direct and semi-direct tests of oral proficiency on the basis of a comparison of test scores alone. Even if candidate scores on the two versions are strongly correlated it may be that the two types of tests tap different language abilities — hence the need for a systematic study of the language actually produced under the two test conditions in addition to a study of test scores.

The current study focuses on one of the key criteria used in Shohamy's (1992) study, lexical density, to analyse candidate output on the direct and semi-direct versions of the **access:** oral interaction sub-test. The findings have implications for the content validity as well as the interchangeability of the two kinds of tests.

3. Lexical density

The term *lexical density* was originally coined by Ure (1971) to provide a measure of the relationship between the number of words with *lexical* as opposed to *grammatical* properties as a percentage of the total number of words in a text. On the basis of her analysis of a wide range of written and spoken texts she found, with only a few exceptions, that the spoken texts had a lexical density of less than 40% (ranging from 24% to 43%) and the written texts a density of greater than 40% (ranging from 36% to 57%). Another focus in this study was on the **presence or absence of feedback** to the speaker in the spoken texts. Texts without feedback (i.e. monologues) all had a density of more than 37% and those with feedback (i.e. dialogues) under 36%. Finally, this research also suggested that **plannedness** may be another important determinant of lexical density with prepared spoken texts all having a lexical density of 37% or higher.

The findings in Ure's (1971) study are important as they suggest lexical density is a valid means of measuring the degree of 'orality' versus 'literacy' in a text, whether written or spoken i.e. texts which are more literate will contain a higher number of lexical words and texts which are more oral will consist of a greater number

of grammatical words. However, her analysis is deficient in that it does not clearly articulate the distinction between words with lexical and grammatical properties and therefore the results should be regarded with some caution.

Halliday (1985) also uses lexical density to compare written and spoken texts in English. Like Ure (1971) he demonstrates that written texts typically contain a higher degree of lexical words than spoken texts. He concludes that the complexity of written language is lexical and that of spoken language is grammatical.

Halliday (1985) proposes two measures of lexical density, the first being identical to the approach formulated by Ure (1971) i.e. the the number of lexical items (where each word is treated as an item) as a proportion of the total number of running words, and the second, which he suggests is a more revealing measure, based on the total number of lexical words as a ratio of the total number of clauses. In the second case, Halliday (1985:80) found that the typical average lexical density for spoken English is between 1.5 and 2, whereas for written English the figure is between 3 and 6, depending on the formality of the writing. This method of calculating lexical density, however, is somewhat problematic as it conflates the problems of determining the criteria for lexical and grammatical items with those inherent in identifying clausal boundaries, an additional difficulty he himself acknowledges:

> *Precisely because it is so fundamental a category, the clause is also impossible to define; nor is there just one right way of describing it.*
>
> Halliday (1985:67)

On the other hand, Halliday (1985) does provide a useful (albeit fairly limited) framework for distinguishing between lexical and grammatical items in a text. Grammatical items are *function* words and operate in *closed*, finite systems in the language. Conversely, lexical items are *content* words and enter into *open* sets which are infinitely extendable. Thus, in English, he suggests, determiners, pronouns, most prepositions, conjunctions and some classes of adverbs are grammatical items. Rather oddly, also included in his initial list are finite verbs but elsewhere, in the examples he uses, these are treated as lexical items. In these examples the verb forms which are consistently classified as grammatical items appear,

appropriately enough, to be modals and auxiliaries as well as all forms of the verbs 'to be' and 'to have'. In addition, all pro-forms (not simply pronouns) and interrogative and negative adverbs are consistently labelled as grammatical. All other adverbs used in the example are treated as lexical items.

Halliday (1985:63) is not prescriptive about this method of classification acknowledging that there is, in fact, a continuum from lexis into grammar. He argues that it does not matter so much where the line is drawn provided it is done consistently. Still, it appears that a detailed taxonomy needs to be devised in order for this analysis to proceed in a principled fashion. One apparent weakness in Halliday's framework is that the division of items under the headings lexical and grammatical is made essentially at the sentence level only. Important discourse phenomena which occur naturally in speech such as discourse markers (words and expressions used to structure discourse including linking and sequencing devices), interjections, (eg *gosh, wow*), reactive tokens (*yes, no, okay* etc.) as well as lexical and non-lexical filled pauses appear to be largely neglected within this system of classification.

Halliday (1985: 64 – 65) does, however, make an important modification to the calculation of lexical density by distinguishing between high and low frequency lexical items. High frequency lexical items are those which occur either commonly in the language in general eg in English *people, thing, way, do, make, get, be, have* and *good* or else more than once in an individual text since repetition reduces the effect of density. In calculating the final lexical density figure the high frequency items are given half the value of the low frequency ones. This would seem to provide a truer, more fine-grained estimate of the overall lexical density.

As Ure's (1971) study suggests, the potential application of lexical density analysis is not restricted simply to contrasts between written and spoken language. Indeed, as Halliday (1985:81) notes, the distinction between speech and writing has become increasingly blurred as a result of modern technology so that oral language samples produced in particular media and/or modes may exhibit greater lexical density than others and likewise for written language samples. It may be that lexical density is a reliable indicator of text type and even text difficulty and therefore has relevance for both language teaching and testing. It could also be

used to analyse the written and spoken output of students and test candidates in relation to registral appropriateness.

In Shohamy's (1992) study it appears that the SOPI produced language which was much more literate than the OPI on the basis of the lexical density figures reported. These results alone strongly suggest that direct and semi-direct tests do not necessarily tap the same kind of language and that therefore their interchangeability is questionable. However, her analysis is perhaps somewhat limited in so far as it takes no account of other important variables which may affect lexical density such as task structure, the amount of preparation and response time and the nature and quantity of verbal or non-verbal feedback provided by the interlocutor in the OPI. It may be that such factors are just as important in determining lexical density as the test format i.e. whether it is live or tape-mediated. In addition, there is a lack of explicitness in relation to whether each word was counted as an item, how precisely lexical and grammatical items were differentiated and whether high and low frequency lexical items were weighted differently for the analysis.

4. Purpose

The current study focuses mainly on the effects of two key variables potentially impacting on candidate output in the Oral Interaction sub-test of the **access:** test — firstly, test format (i.e. live or tape-mediated) and secondly, task type. It addresses this question by examining the degree of lexical density which characterises selected language samples from the two versions of this sub-test. In order to explore these issues the following experimental hypotheses were formulated:

H_{A1}: There is an effect on lexical density for test format.

H_{A2}: There is an effect on lexical density for task type.

H_{A3}: There is an interaction effect on lexical density for test format and task type.

5. The access: Oral Interaction sub-test

The two versions of the access: Oral Interaction sub-test were developed concurrently. The live version was designed to be used with individual candidates and a trained interlocutor in a face-to-face context while the tape-mediated version was designed to be administered to groups of candidates in a language laboratory. A series of matching tasks was developed which were intended to share important characteristics such as task structure and range of expected language functions, although different in content in most cases. A wide variety of tasks were chosen for each version of the test in an attempt to elicit a rich language sample from candidates. The tasks which appeared on each version of the test are shown in Table 1 below.

	Live	Tape
Description	✓	✓
Narration (picture sequence)	✓	✓
Exposition (data based)	✓	✓
Summary (of taped conversation)	–	✓
Discussion	✓	✓
Role play (two way exchange)	✓	–
Role play (answering machine message)	–	✓

Table 1. Task Types

In the first phase of this project, the tests were required to identify candidates at the minimum vocational level of oral proficiency i.e. Level 5[2].

Since the live version of the test was to be conducted overseas with interlocutors who may or may not have been trained language teachers, it was considered essential to reduce the potential

[2]Level five in oral proficiency is described as follows: can communicate effectively in spoken English in a range of social, educational and work situations. Communication is appropriate with a high degree of fluency. Language is grammatically accurate most of the time with a wide range of vocabulary which is used effectively in most situations.

variability in interlocutor behaviour in the direct version to the greatest possible extent. In order to standardise their input as much as possible, a booklet for interlocutors detailing the requirements of the tasks was developed for the live version of the test. The booklet provided them with detailed instructions concerning language input to be used with candidates. In addition, with the exception of the role play, interlocutors were instructed not to intervene once the candidate was clear about the task requirements. This followed from the attempt to provide parity with the semi-direct version of the test. In general, therefore, the live version was designed to be less interactive than direct tests of oral proficiency such as the OPI.

Test booklets were also provided to candidates for both the live and tape-based versions. Both booklets included the stimuli for the test tasks. In addition, in the tape version, the instructions for the tasks were written, as well as spoken on the tape, since the candidate would not have the advantage of clarifying misunderstandings with the interlocutor. The tape version was longer than the live version (45 minutes compared to 30 minutes).

The scoring criteria adopted were identical for matching tasks in the two versions. The criteria used included fluency, grammar, vocabulary, coherence and cohesion, appropriacy of language, intelligibility, overall communicative effectiveness and, in the case of the live version, comprehension. Each of these criteria was assessed on a six-point scale with accompanying descriptors for each level.

The study focuses on the first trialling of the test which was held in Melbourne in December 1992. In this trial 94 candidates attempted both versions of the test. Of these, half were administered the live version first, and half were administered the tape version first to neutralise any potential practice effect. These performances were all audio-taped so that they could be rated retrospectively. Eleven candidates were subsequently excluded because of technical problems with recording equipment. The audio-recordings of both the direct and semi-direct versions of the test from a total of 83 candidates were then rated by thirteen trained raters using the criteria outlined above. Each candidate was independently rated on six occasions.

Multi-faceted Rasch analyses using the programme FACETS (Linacre, 1990) were subsequently run on the test scores from the two versions of the test yielding an overall Pearson correlation for the candidate ability estimates of 0.93 (Wigglesworth and O'Loughlin, 1993). This result suggested a strong relationship between the two versions in terms of candidate performance. The audio-taped recordings then formed the data pool for this study.

6. Method

In order to ensure appropriate sample selection, a stratified random sample of 20 subjects was obtained. This process involved selecting 10 candidates who had completed the live version first and 10 who had completed the tape-mediated version first by drawing candidate numbers at random from each of these two groups. This group of 20 candidates formed an approximately normally distributed range of ability levels using the ability estimates derived from the FACETS programme and, as such, is a reasonably representative sample of the whole cohort of 83 candidates.

Four alternate tasks were chosen as the focus of this study — the description, narration discussion and role play tasks. These were the most directly comparable tasks in terms of task requirements for the candidates. A broad orthographic transcription of the selected language samples was then carried out; a detailed coding of the transcripts was undertaken later as well. This provided a total of 160 language samples for the lexical density analysis. The important features of these matching tasks are shown in *Appendix A: Key characteristics of selected matching tasks.*

In the studies by Ure (1971) and Halliday (1985) it is the *word* (used synonymously with the term *item*) that has been adopted as the basic unit of lexical density (as previously noted, it is unclear whether this is the case in Shohamy's (1992) study). However, while this may be a satisfactory method for an approximate comparison of the relative weight of lexis and grammar in a text, a more refined approach to this analysis would be to focus on the notion of a linguistic *item* as the more appropriate unit of measurement and to differentiate it from the concept of the *word*.

There is, in fact, no one-to-one correspondence between linguistic items and words in English. An item may consist of more than one

word eg multi-word verbs such as *catch up on*, phrasal verbs such as *drop in* and idioms such as *kick the bucket.* Conversely, a word may consist of more than one item eg contractions such as *they're*, and *isn't*. In addition, different items may be realised by the same word eg *lap* : *lap*^{1} (noun or verb as in a race), *lap*^{2} (verb as in 'the cat laps the milk') and *lap*^{3} (noun as in 'sit on my lap'). On the other hand, different words may be realised by the one lexical item (eg *different* and *difference* are alternate word forms of *differ*). Finally, the term 'item' (unlike the term 'word') does not so readily exclude what are sometimes called 'particles' such as *oh* and *mm* which can play important functions (apart from simply expressing hesitation) in spoken discourse. In this study, therefore, it is linguistic items which were counted to measure lexical density in the language samples collected.

A preliminary taxonomy of lexical and grammatical items was drawn up based on a framework devised by Halliday (1985). There were three categories: grammatical items, high frequency lexical items and low frequency lexical items.

The framework was then refined after attempting the analysis on a limited number of the language samples. In order to confirm the viability and robustness of the revised classification system, two independent counts of lexical and grammatical items for three of the candidates on both versions of the test were carried out by myself and a research assistant. The framework was further refined following this stage and then the final version (see Table 2 below) used to analyse the live and tape-based audio-recordings of all 20 candidates.

A. Grammatical items

- Verbs 'to be' and 'to have'. All modals and auxiliaries.
- All determiners including articles, demonstrative and possessive adjectives, quantifiers (eg some, any) and numerals (cardinal and ordinal).
- All pro-forms including pronouns (eg she, they, it, someone, something), pro-verbs, (eg A: Are you coming with us? B: Yes I am), pro-clauses (eg this, that when used to replace whole clauses).
- Interrogative adverbs (eg what, when, how) and negative adverbs (eg not, never).

- All contractions. These were counted as two items (eg they're = they are) since not all NESB speakers regularly or consistently use contractions.
- All prepositions and conjunctions.
- All discourse markers including conjunctions (eg and, but, so), sequencers (eg next, finally), particles (eg oh, well), lexicalised clauses (eg y'know, I mean), meta-talk (eg what I mean, the point is), temporal deictics (eg now, then), spatial deictics (eg here, there) and quantifier phrases (eg anyway, anyhow, whatever).
- All **lexical filled pauses** (eg well, *I mean, so*).
- All **interjections** (eg *gosh, really, oh*).
- All **reactive tokens** (eg *yes, no, O.K., right, mm*).

B. High frequency lexical items

- Very common lexical items as per the list of the 700 most frequently used words in English (accounting for 70% 0f English text) identified in the COBUILD Dictionary project (1987). This list is included in the *Collins COBUILD English Course*, Level 1 Student's book pp 111 – 112 (Willis, J & D, 1988). It includes **nouns** (eg *thing, people*), **adjectives** (eg *good, right*), **verbs** (eg *do, make, get*), **adverbs of time, manner and place** (eg *soon, late, very, so, maybe, also, too, here, there*). No items consisting of more than one word are included in this category as the COBUILD list consists of words not items.
- **Repetition of low frequency lexical items** (see below) including alternate word forms of the same item (eg student/study).

C. Low frequency lexical items

- Lexical items not featuring in the list of 700 most frequently used English words cited above including less commonly used **nouns, adjectives, verbs** including participle and infinitive forms (all multi-word and phrasal verbs count as one item), **adverbs of time, place and manner** and all **idioms** (also counted as one item).

Table 2. Lexical density: classification of items

In carrying out the analysis, all phrasal and multi-word verbs were counted as low frequency lexical items since the COBUILD list of high frequency items only included single words. In addition, only fully audible items were counted. In context, partially or completely inaudible items appear in most cases to have been mispronounced lexical items so that the final lexical density estimates for most of the samples may have been a little lower because of the exclusion of these items. Furthermore, since non-lexical filled pauses (eg *er, um*) were so frequently used by all candidates, they were excluded from the analysis except where they had a clear discourse marking function; it was ultimately considered that their inclusion as grammatical items may have significantly obscured the relationship between lexical and grammatical items in the samplescollected for this study. Finally,where candidates used self-repair, only the final version of an item or utterance figured in the analysis.

The numbers of low and high frequency lexical items and grammatical items in the candidates' output for each of the tasks on both versions were then tallied as frequency counts. The lexical density calculations were subsequently undertaken in two ways following Halliday (1985: 64 – 65). Firstly, no distinction was made between high and low frequency lexical items in calculating the overall lexical density figures — the number of all lexical items was simply expressed as a percentage of the total number of items in each case. Secondly, the high frequency lexical items were given half the weight of the low frequency lexical items, and the weighted number of lexical items were then expressed as a percentage of the total number of items in a given task. Halliday (1985) suggests that this second method represents a more refined approach to determining lexical density. Carrying out the calculations in both ways provided a test of whether it was really necessary to distinguish between high and low frequency lexical items using the weighting system outlined above for this kind of comparative analysis.

For both of these methods the resulting data sets consisted of percentages of the amount of lexicon as opposed to grammar for each candidate as measures of the dependent variable lexical density for each of the eight tasks: live description, live narration, live discussion, live role play, tape description, tape narration, tape discussion and tape role play. Percentages are most safely treated as

ordinal data and the most appropriate measures of central tendency and variability therefore are the median and range respectively. These were calculated for each of the six tasks.
In the study design there were two independent variables: firstly, test format (with two conditions) and task type (with four conditions). The samples were dependent and the data was on an ordinal scale. The experimental hypotheses focused on whether there were significant differences in the degree of lexical density for text format and task type. The most appropriate inferential statistic therefore was a 4 x 2 non-parametric factorial procedure with repeated measures using non-specific hypotheses. The procedure used here is taken from Meddis (1984: 325 – 329).

7. Results

A. Descriptive statistics

Tables 3 and 4 show the median percentage scores, the range of scores, the sum of the ranks and the mean sum of the ranks for each of the eight tasks using both methods of calculating lexical density.

Test format	LIVE				TAPE			
Task	Description	Narration	Discussion	Role play	Description	Narration	Discussion	Role play
median (%)	40.0	38.0	40.0	35.0	42.0	41.0	43.0	41.5
range (%)	31.0 – 49.0	32.0 – 44.0	36.0 – 47.0	31.0 – 40.0	33.0 – 50.0	35.0 – 45.0	37.0 – 47.0	35.0 – 49.0
sum of ranks	93.0	66.0	92.5	36.0	111.0	92.5	121.5	107.5
mean sum of ranks (N=20)	4.7	3.3	4.6	1.8	5.6	4.6	6.1	5.4

Table 3. Method A: Unweighted lexical items (high frequency lexical items assigned the same weight as low frequency lexical items)

Test format	LIVE				TAPE			
Task	Description	Narration	Discussion	Role play	Description	Narration	Discussion	Role play
median (%)	33.5	31.0	32.5	29.0	36.0	34.5	34.5	36.5
range (%)	26.0 – 43.0	26.0 – 36.0	28.0 – 39.0	23.0 – 32.0	29.0 – 44.0	27.0 – 40.0	30.0 – 45.0	27.0 – 44.0
sum of ranks	87.5	65.5	80.5	31.0	121.0	96.0	119.0	119.5
mean sum of ranks (N=20)	4.4	3.3	4.0	1.6	6.1	4.8	6.0	6.0

Table 4 Method B: Weighted lexical items (high frequency lexical items assigned half the weight of low frequency lexical items)

For the first method, where high and low frequency lexical items were not distinguished the median scores across the tasks fall between 35.0 and 43.0 per cent. For the second method, where high frequency items were assigned half the weight of the low frequency ones, the median scores for the eight tasks, not surprisingly, are now lower falling between 29.0 and 36.5 per cent. Graphical representations of these results are useful here in providing an overview of the median scores for each of the two methods.

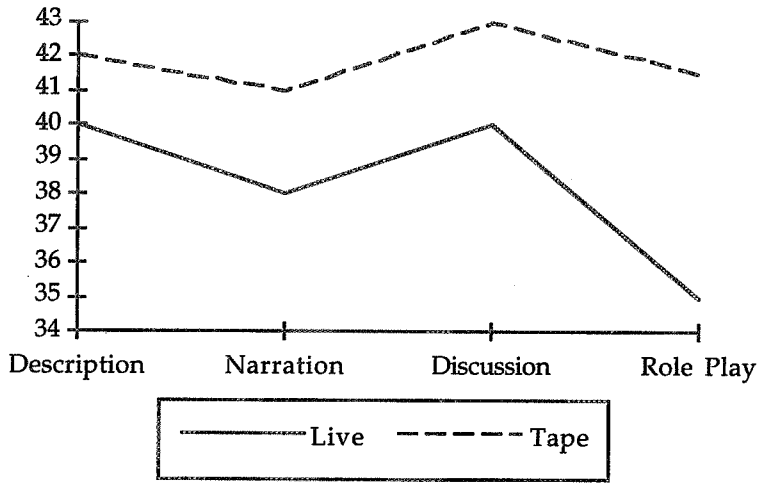

Figure 1. Median scores (%) for lexical density analysis with unweighted lexical items (N = 20).

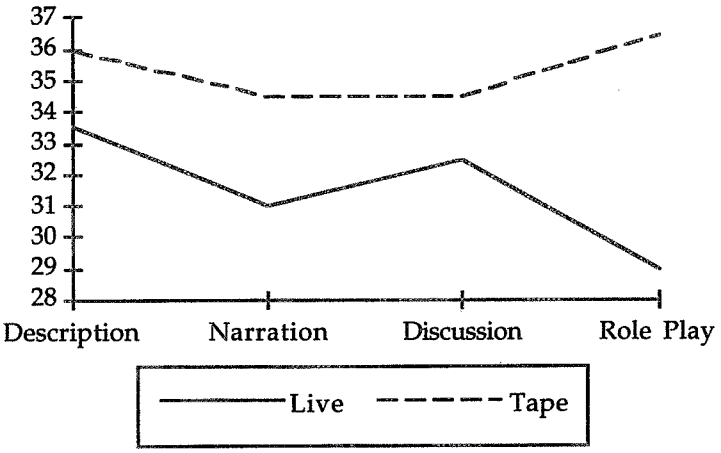

Figure 2. Median scores (%) for lexical density analysis with weighted items.

Looking at the range figures for the first method, the difference between the maximum and minimum scores for each of the eight tasks was fairly broad, from 10 to 18 percentage points. There is a fairly strong disparity between the sums of the ranks — from 36.0 to 121.5. This difference is also reflected in their means which fall between 1.8 and 5.6. For the second method, the difference in the maximum and minimum scores for each of the eight tasks is between 10 and 17 percentage points. Finally, the sums of the ranks fall between 31.0 and 121.0 and their means between 1.6 and 6.1.

B. Inferential statistics

The results of the non-parametric factorial procedure which was used to examine the three experimental hypotheses are reported below.

Method A: Unweighted lexical items

H_{A1}: There is a significant effect on lexical density for test format (H = 21.9, df = 1, $p < 0.01$).

H_{A2}: There is a significant effect on lexical density for task type (H =14.7, df = 3, $p < 0.01$).

H_{A3}: There is no significant effect on lexical density for the interaction between test format and task type (H = 7.2, df = 3, n.s.).

Method B: Weighted lexical items

H_{A1}: There is a significant effect on lexical density for test format (H = 38.0, df = 1, $p < 0.01$).

H_{A2}: There is a significant effect on lexical density for task type (H = 10.0, df =3, $p < 0.05$).

H_{A3}: There is a significant effect on lexical density for the interaction between test format and task type (H = 9.4, df = 3, $p < 0.05$).

For the two methods of calculating lexical density the results are similar but not identical. In both cases the effect for test format is significant at the 0.01 level, while the effect for task type is significant at the 0.01 level and at the 0.05 level respectively. In statistical terms, the most important difference occurs in the results for the interaction effect between text format and task type. The result is not significant using the first method but significant at the 0.05 level when the second method was employed.

The findings using the more finely-tuned second method of determining lexical density (where high frequency lexical items were assigned half the weight of low frequency items) are probably the more accurate here. The slight discrepancy in the results based on the two methods suggests that this more refined analysis is probably warranted in formal investigations of lexical density, especially where inferential statistical procedures are to be employed.

In any event, all of the findings reported above require further interpretation as they provide no real indication about the nature or size of the particular effect, even where it is significant. In the absence of a suitable *post hoc* procedure, a visual representation of the mean sums of the ranks for the two methods of determining lexical density makes it possible to clarify the statistical results more fully.

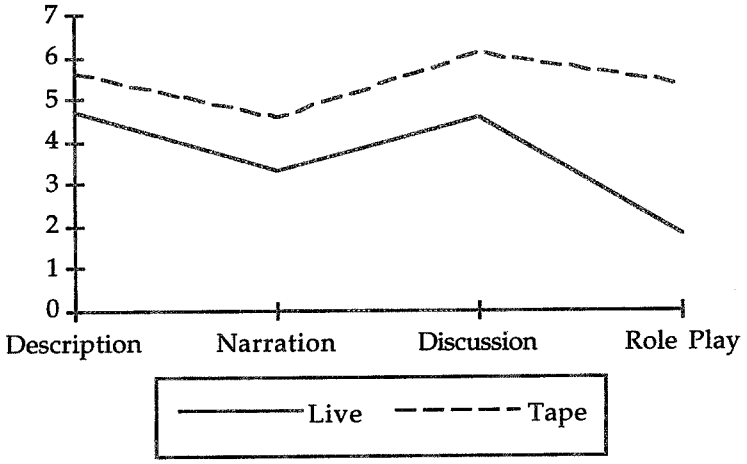

Figure 3. Sample rank means for lexical density analysis with unweighted lexical items (N=20).

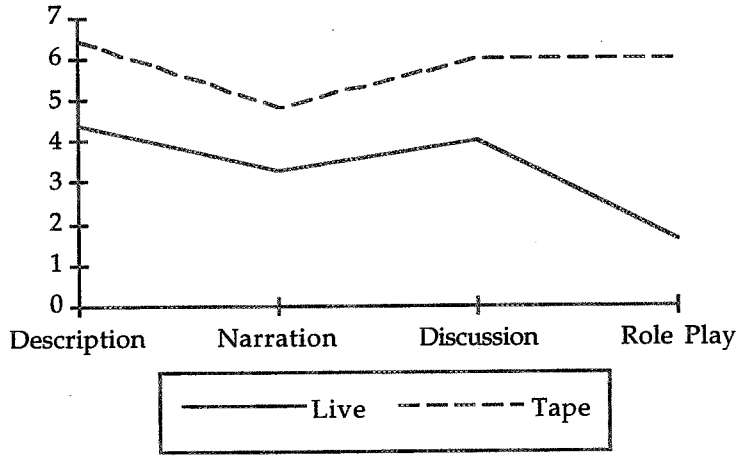

Figure 4. Sample rank means for lexical density analysis with weighted lexical items (N=20).

Clearly, in both graphs the lexical density in candidate output on the tape-based version is higher for all four tasks than on the live version. In addition, it appears that the degree of lexical density

was lower for the narration task than the description and discussion tasks for both versions. In the case of the role play, the degree of lexical density was similar to the description and discussion tasks in the tape version but clearly lower on the live version than all of the other three tasks. In relation to the third hypothesis, the fact that in both graphs the two lines run almost perfectly parallel for the first three tasks and then diverge on the role play task suggests that any real interaction effect overall between test format and tasks type stems from the larger difference in lexical density on this last task compared to the other three tasks. The fact that the interaction effect was not statistically signficant when the high and low frequency lexical items were unweighted in calculating the lexical density but significant when weighted, appears to be simply a function of the size of the difference between the sums of the ranks for the live and tape role play tasks in each case. This difference was slightly greater using the weighted method, hence the result obtained for the interaction effect reaches the 0.05 significance level.

8. Discussion

Although the results indicated that the effects of test format and task type as well as the interaction effect between them on the lexical density in candidate output were statistically significant, the differences for either of these two variables do not appear to be large overall. This is true particularly when compared, for example, with the findings attributed to the effect of test format in Shohamy's (1992) study i.e. 40% lexical density on the OPI and 60% on the SOPI. The range of median percentage scores across all eight tasks was quite narrow for both methods of calculating lexical density, only 8 and 7.5 percentage points respectively.

As suggested in both the graphical representations of the median percentage scores (Figures 1 and 2) and the the sample rank means (Figures 3 and 4), the most salient difference which emerges from this study in terms of the effect of either test format or task type on the degree of lexical density is between the live and tape role plays. A possible explanation for this finding relates to the amount of feedback given to candidates in each of the eight tasks examined in this study.

In the description, narration and discussion tasks on the live version, the interlocutor feedback was extremely limited in all cases, normally consisting only of reactive tokens such as *mm, yes* and *right.* This followed from the instructions to interlocutors not to actively intervene once the candidate was clear about the requirements for these tasks. An important component of the live role play, however, was that interlocutors were required to make a substantial contribution to the interaction, their input throughout the conversation playing a crucial role in shaping the content of the candidate's output. The fact that the median percentage score is clearly lower on the live roleplay than the other three tasks using either the unweighted or weighted method for calculating lexical density suggests that the nature of the feedback in any given task will strongly influence the degree of lexical density i.e. the higher the level of feedback from the interlocutor the lower the degree of lexical density in candidate output. This conclusion is supported by examining the tape results as well. A distinguishing feature of all tasks on the tape version is the total absence of feedback to the candidate. The median percentage scores for all four tape tasks are higher than for any of the live tasks.

On the basis of the results for the effect of task type it appears that task structure may also impact on candidate output in tests of oral proficiency i.e. 'open' tasks seem to elicit language with a higher degree of lexical density than 'closed' tasks (except for the live role play for reasons outline above). In either case the candidate's response will only be as lexically dense as the task allows. In each of the narration tasks the stories to be told (using the sequence of pictures) were fairly simple with little room for interpretation, potentially limiting the use of the candidate's lexical resources. Perhaps the more challenging requirement of these tasks is relating the pictures appropriately to each other using discourse markers, pro-forms and other cohesive devices — all grammatical items. This may account for the relatively lower levels of lexical density in both narration tasks. By contrast, in the more open tasks — notably, the description and interview — candidates are not constrained by any stimulus material and are therefore likely to display a greater range of their lexical resources. Hence, the higher degree of lexical density in these cases. However, again, as the differences in the median scores between tasks is not great, such an interpretation is only offered tentatively. Still, it does suggest an avenue for further research.

Although not addressed in the experimental hypotheses there are other factors which may influence the degree of lexical density in the samples collected. One of these is whether or not candidates planned their responses i.e. planned answers are likely to be more lexically dense than unplanned ones. If this was the case then there should be a greater difference between the two description tasks — the live one which did not include planning time and the tape-based one which did — than either of the other tasks which had provision for preparation on both versions of the test. But examination of the median scores and the sum of the ranks for both methods of calculating lexical density does not yield a clear answer to this question. This is a problematic variable to investigate, however, since, even where planning time is provided, there is no guarantee it will be effectively used by the candidate. In addition, where planning time was built into the design of the task in the live version, candidates often failed to use the full amount of time allocated for this purpose, either through their own choice or because the interlocutor had cut short the preparation time.

Next, the amount of response time allowed may also have an effect on lexical density. It is possible that one of the reasons why the tape-based language samples are generally more lexically dense is that candidates are conscious of the limited time frame in which they must speak before and during their performance and tailor their communication accordingly. By contrast, the response time in the live version is not fixed in advance and so the candidate may feel under less pressure to include a maximum amount of content in a short space of time. An interesting way of exploring this issue might be to compare levels of lexical density on the tape-based version by varying the response time for similar tasks.

One other variable which could influence lexical density is candidates' perception of when their performance will be assessed. In tape-based tests it is obvious that this will occur later in time so that candidates are clear that their communicative goal is to create a record of their performance for raters displaced in time and space. In live tests, however, it is not always apparent when the assessment will occur. It is possible in the live version of this test that candidates assumed the assessment was being carried out at the time of the test, especially given the presence of an observer as well as the interlocutor, even though they were aware that their performance was being recorded. Where this was true it is likely to

have lead to a difference in the candidate's focus in the two versions, that is, a greater orientation towards content or product in the tape version and interaction or process in the live version. A stronger focus on content is, in turn, likely to result in higher lexical density while the reverse is likely to be true for a stronger orientation towards interaction. Where candidates perceived the assessment as occurring later in time they may have been more focused on product as in the tape version.

It is, of course, impossible to separate candidates on this basis with any degree of certainty retrospectively. However, this interpretation might account for why the lexical density figures for a minority of candidates were as high on the live version as the tape version i.e. these candidates correctly assumed that the assessment of their performance would be carried out later using the audio-recording as in the tape version. Greater parity in lexical density between the two versions may be achieved if candidates are made aware in advance that their performance will be rated at a later time in both instances.

9. Conclusion

The investigation into lexical density on the four selected matching tasks on the Oral Interaction sub-test of the **access:** test suggests that the tape-based version taps a slightly more literate kind of language than the live version. However, this difference is probably not of a sufficent magnitude to threaten the potential interchangeability of the two versions overall. It should be stressed that there are probably a number of factors including task structure and interlocutor feedback as well as preparation and response time which influence candidate output in oral proficiency tests, not simply whether the candidate is talking to another person or a microphone *per se.*

Interlocutor feedback, rather than test format, emerges as perhaps the single most important determinant of candidate output in this study. The results indicated that the higher the level of feedback from the interlocutor the lower the degree of lexical density in this output. If this is the case, then altering the live version of the **access:** test to make all of the tasks more interactive could well result in the two versions tapping language which is qualitatively even more different than is the case here. Greater interactiveness

may therefore enhance the live version's content validity as a speaking test since conversation is the primary domain of oral communication but, at the same time, reduce its potential interchangeability with the tape-based version.

References

Halliday, M.A.K. (1985) *Spoken and written language.* Melbourne: Deakin University Press.

Linacre, J.M. (1990) *FACETS: computer program for many-facteed Rasch measurement* (version 2.62). Chicago, IL: Mesa Press.

Meddis, R. (1984) *Statistics Using Ranks.* New York, NY: Basil Blackwell Inc.

Shohamy, E. (1992) The validity of concurrent validity of direct versus semi-direct tests of oral proficiency. *Paper presented at the AERA Convention, San Francisco, CA, April 1992.*

Stansfield, C. (1991) A comparative analysis of simulated and direct oral proficiency interviews. In S. Anivan (ed) *Current developments in language testing.* Singapore: RELC.

Stansfield, C. and D. Kenyon (1992) Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System* 20.3. 347 – 364.

Ure, J. (1971) Lexical density and register differentiation. In G.E. Perren and J.L.M. Trim (eds) *Applications of linguistics.* Cambridge: Cambridge University Press.

Wigglesworth, G. and K. O'Loughlin (1993) An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English. *Melbourne Papers in Language Testing* 2,1: 56 – 67.

Willis, J & D (1988) *Collins COBUILD English Course, Level 1, Student's book.* London: Collins.