Examining the validity of two instruments as measures of curriculum-related L2 vocabulary development

Anne Isaac
The University of Melbourne

Abstract

The exploratory study described in this paper attempted to gauge the validity and usefulness of two instruments as measures of curriculum-related vocabulary development. The study involved five non-native speaker (NNS) subjects and twenty native speaker (NS) subjects. The NNS subjects were from a variety of language backgrounds and had completed a 48-hour English for Academic Purposes (EAP) course which employed a stylistic approach to literature. In post-course evaluations, the NNS subjects attributed improvements in vocabulary size and in their knowledge of word choice to the course they had taken.

Instruments chosen were the Lexical Frequency Profile (Laufer and Nation, 1995) considered to be a reliable measure of lexical size/ richness, and a multiple choice (MC) cloze task devised by the researcher to target the subjects' knowledge of collocations, semantic appropriateness, and sensitivity to style. Baseline data for those cloze items designed to elicit collocational knowledge was provided by the NS subjects who also completed the MC cloze task.

Data included the NNS subjects' results on pre-course placement tests; two compositions—one written by the subjects as part of the pre-course placement test, the other written post-course—each analysed using the Lexical Frequency Profile (LFP); their responses to the receptive MC cloze and their audio-taped comments made during an immediately retrospective interview; the NS subjects' responses to the cloze task; and audio-taped interviews with the NNS subjects held eight months after the end of the EAP course.

Results of the data analysis firstly revealed the strengths and weaknesses of the LFP as a longitudinal measure of vocabulary size/richness and indicated the need for modifications in standard procedures if the benefits of the test are to be maximised. Secondly, analysis of the NNS and NS subjects' responses to the MC cloze

showed that this type of task provided some measure of collocational knowledge, but lacked validity as a test of stylistic awareness.

Despite its small scale and limitations, the study provides some practical directions for further research into ways of measuring curriculum-related vocabulary acquisition.

Introduction

Vocabulary acquisition, often considered to be central to learning a second language (L2) (Laufer, 1997; Sökmen, 1997; Lewis, 1993), is also a relatively neglected area in applied linguistic and language testing research. The foremost challenge for test designers and teachers alike is to define what 'knowing a word' means. In addition, for test designers, it is to construct reliable, and particularly integrative, measures of the kind of lexical knowledge required for communicative purposes (Read, 1997); for teachers, the challenge is to know what to teach, to develop pedagogies to facilitate vocabulary learning, and to know how best to test curriculum-related vocabulary development.

Context of the present study

The study described in this paper was an attempt to examine the validity and the usefulness of two instruments as measures of curriculum-related vocabulary acquisition in students for whom English is a second language (L2). The instruments, the Lexical Richness Profile (Laufer and Nation, 1985) and a test of collocational knowledge and stylistic awareness, were used to measure the vocabulary development of L2 university students who had taken a 48-hour, content-based English for Academic Purposes (EAP) subject as part of their degree. The semester-long subject, entitled *Language Approaches to Australian Literature* (LAAL) employs a stylistic approach to a variety of texts from Australian literature and focuses on the development of the participants' reading and writing skills and vocabulary knowledge.

The research presented in this paper was motivated by the findings of a previous evaluation study (Isaac, 1998), and by the post-course evaluations of the participants in the present study. The post-course evaluations were completed in written form at the end of the LAAL course, and eight months later, in an interview with the researcher.

The evidence from both cohorts of LAAL students showed clearly that they perceived vocabulary development to be the most obvious learning outcome of the course. The students reported some increase in vocabulary size, but more importantly, an improvement in the quality of their vocabulary knowledge. They mentioned having a greater awareness of the appropriate use of words in specific contexts; consulting a monolingual dictionary and a thesaurus more frequently to find the 'right' word; and a decreased dependence on bilingual dictionaries.

The need to find objective evidence in support of the students' perceptions led to an examination of vocabulary measures which might be employed to test the types of vocabulary development cited by the students.

Vocabulary knowledge: definitions and dimensions

Definitions of what it means to know a word vary, but most researchers agree that vocabulary knowledge is a matter of degrees of familiarity, from partial to precise, and that knowledge about a word develops in stages. Multiple knowledge frameworks, proposed by Nation (1990), Melka (1997), Laufer (1997) and Laufer and Paribakht (1998) among others, include several degrees of familiarity: having visual, phonological, morphological, syntactical and lexical information about a word, and knowing when its use is situationally or stylistically appropriate.

Higher degrees of familiarity are generally taken to include knowledge of the multiple meanings of words (polysemy), the stylistic qualities of words, and collocations, combinations of lexical or grammatical words which co-occur with great frequency and exclusiveness (Schmitt, 1998a). According to Schmitt, the adjective 'blonde', for example, is almost always in combination with 'hair'. Applied linguists tend to agree on the importance of collocations in L2 language learning (Bahns and Eldaw. 1993; Carter, 1998; Schmitt, 1998a), but indicate that collocations are difficult to measure, teach and learn. For linguists the problem is to know what degree of frequency is required before a particular combination of words may be called a collocation.

Frameworks which describe or measure lexical knowledge refer to certain dimensions of knowledge: passive/receptive and active/

productive; breadth/size and depth/quality. A continuum of knowledge is generally agreed to separate word recognition and comprehension (receptive knowledge) on the one hand, and word use (productive knowledge) on the other. According to Melka (1997), movement from one stage to the next is imperceptible, making it difficult to identify the point at which receptive knowledge becomes productive.

Vocabulary size has traditionally been considered as the most important indicator of word knowledge, although attempts to correlate vocabulary size with reading comprehension have been inconsistent (Nation, 1990). Depth of vocabulary knowledge refers to the multiple aspects of word knowledge described in the theoretical models of vocabulary knowledge outlined above.

Testing vocabulary knowledge

Discrete tests

The majority of existing lexical measures test word knowledge as an independent construct. Tests of size and depth outlined below are all in this category.

Tests of size

Studies of the vocabulary breadth of non-native speakers (NNS) have mostly been based on samples taken from word frequency lists, or to a lesser extent, from dictionaries. Two well-known measures of L2 passive vocabulary size are the Eurocentres' Vocabulary Size Test (EVST) (Meara and Jones, 1990) and Nation's Vocabulary Levels Test (1990), both based on word lists from Thorndike and Loge (1944).

Wesche and Paribakht (1996) argue that the accuracy of these breadth measures is dependent on the representativeness of the corpus used and its compatibility with the individuals tested. (p.16). Hence words may appear at different frequency levels in different corpora. Furthermore, these measures test the respondents' receptive knowledge, relying on recognition and self-reporting about known/unknown or real or invented words on a checklist (the EVST is an example of this type); or on the respondents' ability to match target words with synonyms or short analytical definitions in multiple-choice format (cf. Nation's Levels Test).

Other discrete tests have been used to measure the active size and richness or variation of vocabulary as it is used in L2 writing. Lexical richness measures include lexical density (the percentage of lexical as opposed to grammatical words in a text); lexical originality (the percentage of words used only by one particular writer in a cohort); lexical sophistication (the percentage of less frequent or 'more advanced' words in a text) and lexical variation (the ratio between the number of word families or types and the total number of word tokens used in a text). A more recently developed measure of lexical richness, the Lexical Frequency Profile (Laufer and Nation, 1995) classifies the total number of word types used according to the frequency level to which each word family belongs. A number of studies cited by Laufer and Nation (1995) and Laufer (1998) have indicated a relationship between writing proficiency and lexical richness (e.g. Astika, 1993; Engber, 1993; Linnarud, 1986).

Vocabulary size tests are of limited value in assessments of curriculum-related vocabulary learning. Their major shortcoming, according to Read (1997) and Wesche and Paribakht (1996), is that they show neither degrees of knowledge of the words 'known' by the test-taker, nor the stages by which that knowledge has been acquired, nor "which knowledge is enough" (Wesche and Paribakht, 1996: 27) to function effectively. In addition, vocabulary size tests do not measure knowledge of words with multiple meanings and uses (Read, 1993).

Tests of L2 vocabulary depth

Depth is a more difficult dimension of vocabulary knowledge to measure (Schmitt and McCarthy, 1997), partly because of practical considerations, such as the issue of how to include an adequate sample without devising a test which takes too long to administer in either written or oral format. As a result, relatively few tests have been designed to measure L2 vocabulary.

Read (1997) distinguishes between two approaches to measuring depth of knowledge: the developmental approach, characterised by the use of a scale describing the stages of acquisition of a word; and the dimension approach which describes the degree to which different types of word knowledge are receptively and productively known.

The Vocabulary Knowledge Scale (Wesche and Paribakht, 1996) and the Word Associates Test (Read, 1993; 1995) are two tests which have met the criteria of reliability and practicalness mentioned above and which exemplify the developmental and dimensional approaches respectively .

The Vocabulary Knowledge Scale (VKS) consists of a five-point scale designed to reflect/test stages of acquisition of the core knowledge of given words. The stages of knowledge tested are whether:

(1) the word is not familiar;

(2) the word is familiar but its meaning is not known;

(3) a correct synonym or translation can be provided;

(4) the word can be used with semantic appropriateness in a sentence; and

(5) the word can be used with semantic appropriateness and grammatical accuracy in a sentence.

The VKS combines self report (Stages 1 and 2) and performance tasks (Stages 3-5) and tests passive as well as active knowledge. Its effectiveness in measuring relatively small changes in vocabulary development over a short instructional period has been demonstrated. However, Read (1997) has raised doubts about the validity of the equal-interval scale, particularly given the gap between word recognition (Stage 2) and evidence of the word's analytic or paradigmatic relations (Stage 3). Likewise, for Schmitt (1998b), the inherent weakness of developmental tests such as the VKS lies in attempting to locate and number stages in a process (acquisition) which is considered to be a continuum.

The Word Associates Test, an example of the dimensional approach, measures the learners passive knowledge of semantic (syntagmatic, collocational and analytical) relationships between words derived from the University Word List developed by Xue and Nation (1984). For each of the 40 items tested, the learner is given eight words: four are distractors and the other four synonyms, collocates or parts of the definition of the stimulus word. While the test has been shown to be valid and practical, Read himself has questioned the role of guessing in the test outcomes (1997).

Aside from these tests of depth, instruments have been devised to measure partial word knowledge (Joe, Nation and Newton, 1998) or multiple aspects of vocabulary knowledge. Dimensional measures based on a multiple knowledge framework have been used by Viberg (1993), Schmitt and Meara (1997) and Schmitt (1998a; 1998b; 1998c; 1999a; 1999b) to investigate the acquisition and use of a limited number of vocabulary items – 20 nuclear verbs in Viberg's study; as few as 6 items in some of Schmitt's research (1999a). Schmitt has designed experimental procedures to measure higher degrees of word knowledge, such as collocations (1998a; 1999b), intuitions of word frequency (1999a) and responses to word associations (1998b). Other studies by Schmitt (1998c; 1999b) and Schmitt and Meara (1997) have used a word knowledge framework approach to study the interrelationship of multiple aspects of lexical knowledge, including size, spelling, word associations, grammatical information, collocational knowledge and polysemy.

Apart from these studies, the researcher found no references in the literature to measures of sensitivity to style/register except for a suggestion by Oller to extend the VKS to reflect a sixth level of knowledge to take into account polysemy and different contexts of use (cited in Wesche and Paribakht, 1996).

Integrative measures

In a comprehensive overview of vocabulary testing (1997) and in a more recent paper (1999), Read argues for the use of context-dependent tasks which measure the test-taker's ability to comprehend or use vocabulary with reference to a larger linguistic context. Equally, he advocates the assessment of vocabulary as one component of integrative testing of specific language skills.

Read (1997) and Wesche and Paribakht (1996) mention four main types of integrative measures: those which assess knowledge of specific words in reading comprehension tests (for example as in the case of the TOEFL Test after 1995); dictation tests; writing and speaking tests assessed by reference to multiple, descriptive rating scales, including a vocabulary scale; and cloze tests.

Despite long-standing debate, the cloze procedure is assumed by many researchers to provide a measure of lexical knowledge. Research by Jonz (1990) and others has shown that around 42% of

items on fixed ratio cloze tests measure lexical knowledge, while 34% of items were related to textual cohesion. Carter (1988) has demonstrated the key role of cloze in sensitising students to the presence of lexical chains within texts and thus to elements of textual cohesion.

Rational deletion cloze tests can be used to an even greater extent than the standard random deletion cloze to focus on assessment of lexical knowledge. Bahns and Eldaw (1993) combined a translation task and a rational deletion cloze task to test students' receptive knowledge of 15 collocations. No significant difference was noted in the results on each task.

Several variants of the basic cloze procedure featuring either random or rational deletion have been used as a measure of L2 vocabulary. The C-Test, which requires the test-taker to complete the second half of every second word, is one such variant. Chappelle's research (1994) into the validity of the C-test indicated that while there was a correlation between vocabulary knowledge and test results, performance on the test was also dependent on a number of factors, including the exact words and word types of the partially deleted items. Another variant is the multiple choice (MC) cloze, where the test-taker is provided with a limited set of choices—of which only one is correct—to fill each gap.

In the *stylistics variants cloze*, which is related to the MC cloze procedure, the choices are similar, but differ in acceptability by their style or register (Soudek and Soudek, 1983). The stylistics variants cloze is presented by Soudek and Soudek (1983) and by Carter (1988) as a rational deletion procedure for use in teaching rather than testing.

The examination of tests of size and depth conducted by the researcher underscores the fact that owing to the many dimensions of word knowledge, all vocabulary tests have their limitations. Equally, it highlights the need for a multiple test approach when trying to measure aspects of lexical knowledge.

Purpose

The aim of the present study was to examine the sensitivity with which two instruments might be able to capture the vocabulary

development of the LAAL students over a relatively short time span. Specifically, the research questions focused on

* the validity of the Lexical Frequency Profile (LFP) as a discrete measure of increase in the size/richness of the students' free active vocabulary knowledge; and

* the validity of a multiple choice (MC) cloze as a context-dependent measure of higher degrees of receptive word knowledge, including collocational knowledge, semantic appropriateness and stylistic awareness.

These measures were chosen over other tests of size and depth mentioned in the overview for a variety of reasons. Unlike other measures of size, such as the EVST or the Levels Test, which would have required repeated applications before and after the LAAL course, the LFP enabled the researcher to analyse the L2 subjects' vocabulary production in their writing in the pre-course placement test by comparison with that in a post-course composition. The LFP has been shown to be a valid and reliable measure of lexical use in free written production, and one which can discriminate between learners placed at different levels of proficiency (Laufer and Nation, 1995). Laufer and Nation (1995) also found the LFP measure to be topic-independent, providing the topics used are of a general nature. This characteristic provided particular impetus for choosing the LFP measure, since two placement tests are used (one for international NNS students, the other for local NNS students) in the university where the study was based. Both placement procedures require the test-taker to write a composition of a similar length and on a similar, controversial topic, in a one-hour session.

The MC cloze test, on the other hand, was chosen over the Word Associates Test and the VKS because it offered the possibility of measuring higher orders of word knowledge, including stylistic awareness, a key focus of the vocabulary component of the LAAL curriculum. The test had originally been designed to be taken by the experimental subjects as well as by a control group of L2 subjects with comparable backgrounds and proficiency levels, thus eliminating the need for a pre-treatment test. Although difficulty in recruiting an adequately sized control group made it impossible to pursue this objective, the MC cloze was retained.

Methodology

Given the exploratory nature of the project, a small subject pool was recruited and a more in-depth qualitative approach was adopted. Schmitt and Meara (1997) comment on the value of looking at individual behaviour in depth of knowledge studies: "important information about individual behaviour is often lost in group analyses" (p. 20).

Subjects

Two groups of subjects participated in the present study: one was a group of NNSs, and the other a group of native speaker (NS) subjects.

NNS subjects

All students enrolled in LAAL were invited to participate in the project towards the end of the course. The five subjects, two males (G and P) and three females (J, M, and X), were recruited on the basis of their availability and expressed interest in the project.

The subjects came from a variety of language backgrounds, were reasonably alike in terms of their length of residence in Australia and with the exception of P and X, were close in age (Table 1). The subjects were representative of the students enrolled in LAAL in terms of their resident status, academic status, and gender. Their language proficiency levels were also representative of the total student cohort. All subjects had intermediate to advanced levels of language proficiency judging by their results on two pre-course local placement tests (Figure 1). According to the test results (converted to a percentage), P was ranked first in terms of overall (reading and writing) proficiency, followed by M, G, J and X. Nevertheless J and X had the strongest reading proficiency, followed by P, M and G. In terms of writing scores, P was most proficient with a score of around 88%, followed by G and M with scores of 77%, and X and J with scores of around 66%. P and G shared equal sub-scores on Vocabulary and Grammar (77%) as did M and X (70%), followed by J (66%). The range of language competence in reading and writing skills at the start of the course was therefore both varied and broad.

Subject	G	J	M	P	X
Language background	Japanese	Indonesian	Mandarin	Mandarin	Cantonese
Age	20	24	20	36	33
Gender	M	F	F	M	F
Resident status *	I	I	P	I	P
Time in Australia	> 4 yrs	< 6 months	> 1 yr	< 1 yr	> 1 yr
Academic status & course	UG Arts /Science	PG Science	UG Arts	PG Education	UG Engineering /Computer Science

* I = international; P = permanent resident

Table 1: NNS subject profile

Moreover, there were a number of differences relating to the length and degree of the subjects' exposure to English: G, P and X had sat their diagnostic assessments five months prior to the beginning of the LAAL course, whereas J and M had arrived in Australia and completed their diagnostic assessments only one or two weeks before the start of the course. This meant that G, P and X had been exposed to an English speaking environment for twice as long as their fellow NNS subjects by the time they wrote the post-course assessment. Further differences included the fact that X had completed a 60-hour EAP course between taking the placement test and enrolling in LAAL, while G was studying LAAL and the 60-hour EAP course concurrently.

NS subjects

A second group of subjects were native speakers who completed the same cloze task as that completed by the NNS subjects, thus providing a means, albeit crude, of validating the researcher's intuitions about items designed to elicit collocational knowledge. For Schmitt (1998a), counting co-occurrences of words appearing in large corpora is more objective and thus preferable to obtaining NS intuitions as a way of collecting norming data for collocations.

However, given that the multiple choice alternatives to the cloze task sometimes involved fine distinctions, it was considered more appropriate to collect the responses of local NS subjects on the identical task completed by the NNS subjects. The NS group consisted of 10 males and 10 females whose ages were between 19 and 55. Four NS subjects were university students in a range of disciplines (languages, politics and computer science), two were graduates (in health sciences and social work), eight were academics (in applied linguistics, mathematics and science), two were English teachers, and four were editors.

Instruments

Lexical Frequency Profile

The LFP entails the application of a computer program, called VocabProfile, which matches vocabulary used by subjects in free productive writing, like a composition, with words contained in three word lists: the first 1000 most frequent words, the second 1000 words and the University Word List (UWL), containing 836 word families beyond the 2000 word level and occurring frequently across a range of cross-disciplinary academic texts. The resulting frequency profile indicates the proportion of word families from different frequency levels as well as those not contained in any of the lists. According to Laufer (1995), the LFP analysis of the vocabulary of more advanced students can also be shown as a condensed profile made up of the percentage of the first 2000 words as compared with the percentage of words in the UWL and words not in any of the other lists (NIOL). A second condensed method mentioned in Laufer and Nation (1995), leaves aside the 1000 most frequent words, and profiles only words in the second 1000, the UWL and NIOL words.

MC Cloze

The second test used, a multiple choice cloze, was designed by the researcher to measure various aspects of the depth of the subjects' vocabulary knowledge, including collocational knowledge, stylistic awareness, and semantic appropriateness (Appendix 1).

The MC cloze task was based on a short story extract describing the experiences of a young Chinese woman raised by an Indian family in the predominantly Anglo-Saxon environment of the Australian

federal capital, Canberra. The text was chosen because it dealt with a topic with which the subjects were likely to be able to identify, given their own backgrounds, and because the language used seemed sufficiently colloquial to be comprehensible to the subjects.

In view of the experimental nature of the project, a restricted number of items (seven in total) was used, considerably fewer than would be customary in a conventional MC cloze. Three items (3, 6, & 7) were selected to test knowledge of collocations and the remaining four tested awareness of semantic appropriateness and/or style/register (1, 4 and 5), or both collocations and style (2), and in one case, grammatical accuracy (5). Three synonyms or similar words were provided for each deletion, although the alternatives were distinguished by semantic or stylistic differences in usage.

A Lexical Frequency Profile analysis of this passage using the first condensed method showed that almost 78% of the vocabulary was at or below the 2000 level, and 22% was beyond 2000. This indicated that although the text would be challenging for intermediate and advanced L2 students, it was not likely to be beyond their proficiency levels. Nevertheless, the analysis also showed that while two of the deleted items (Item 2, *stock*; and Item 7, *sea*) were at the 1000 level, all the other items were in the NIOL category and would thus pose some difficulty for some subjects.

Scoring of the cloze test (shown in Appendix 1) reflected trends in the NS responses (Figure 2). Scores on each item ranged from 0 for a semantically or grammatically inappropriate or unlikely response, to 1 for a semantically appropriate but stylistically inappropriate or unlikely response, to 2 for responses that were semantically and stylistically appropriate or reflected collocational knowledge. In the case of Item 5, the alternatives *continued* and *lingered* were given an equal score of 2, since the balance of NS responses favoured *continued* over *lingered*.

Data

Data collected comprised the NNS subjects' results on the placement test they completed prior to the course; their performances on two free productive writing tasks, one a composition written as part of the pre-course placement test, the other a post-course composition; their responses to the receptive cloze task which were complemented by an

immediately retrospective audio-taped interview; the responses of NS subjects to the cloze task; and finally, comments made by the subjects during the post-course audio-taped interview held eight months after the end of the LAAL course.

Procedures

Five to six weeks after the end of the LAAL course, the five NNS subjects were asked to write an argumentative composition of approximately 300 to 400 words within a one-hour time frame on the following topic:

The advantages of studying in a second language/ culture outweigh the disadvantages.

The post-course compositions were compared with the subjects' pre-course placement compositions, completed under identical conditions. As mentioned above, the use of two pre-course assessment procedures meant the pre-course compositions were on one of two topics, each similar to one another and to the post-course composition.

The pre- and post- course compositions were edited, following Laufer and Nation (1995), and entered into a computer loaded with the VocabProfile program. Editing included pruning each subjects' two compositions to an identical length (the last words in the longer composition were deleted); deletion of proper nouns, incorrect use of lexical items and any wording borrowed from the composition prompts; and correction of spelling errors. The samples were analysed with the LFP measure and, given the intermediate to advanced proficiency levels of the subjects, the frequency profiles were subsequently converted according to the two condensed profile methods described above: namely, the 'beyond 2000' profile made up of the percentage of the first 2000 words as compared with the percentage of words beyond 2000; and secondly, the condensed profile made up of words in the second 1000 words, the UWL and the NIOL words.

The cloze test was completed by the NNS subjects on the same occasion as the post-course composition. To eliminate any bias due to order, the cloze was completed directly before the composition in the case of two subjects, and directly afterwards in the other three cases. The completion of the cloze task was accompanied by the subjects'

immediately retrospective explanations of the rationale governing their selection from the alternatives offered. These interviews were audio-taped.

The cloze task was also completed by the 20 NS subjects, followed by a retrospective commentary, albeit briefer than that of the NNS subjects, and documented in note form.

Results

LFP analysis

Figure 3 illustrates results of the analysis following the first condensed profile method, which contrasted words in the first 2000 lists and those beyond 2000. The subjects' results as a group indicated that the majority of vocabulary used in the placement test composition and in the post-course compositions was at the 2000 level or below. In terms of individual results, the comparative analysis of vocabulary used in each composition indicated that noticeable changes took place in the case of P, and to a lesser extent, X, who had both taken the placement test five months prior to enrolling in LAAL. The beyond 2000 vocabulary of P had increased by approximately 9%, with a corresponding 9% reduction in words in the first 2000 levels. Similarly, X demonstrated a 6.5% increase in words beyond the 2000 level in her post-course composition. On the other hand, the results of G, who had also taken a placement test one semester prior to taking the LAAL course, showed a decrease of nearly 2% at the beyond 2000 level, and a corresponding increase in words in the first 2000 lists. There appeared to be only a very slight change for the subjects who took the placement test and LAAL in the same semester: J's beyond 2000 level vocabulary increased by 1.83%, while that of M decreased by 1.44%.

Results produced by the second condensed method (profiling vocabulary in the three categories beyond the first 1000 words) indicate that some vocabulary enrichment occurred at the UWL level in the cases of P , X and J (Figure 4). P's profile showed an increase of nearly 14% at the UWL level; X used 4.7% more words at this level, and 29.7% more words in the NIOL level by comparison with her placement test writing; J's use of words increased by 3% at the UWL level and by 11% at the NIOL level. In each of these three cases there was a corresponding decrease in the presence of vocabulary at the

second 1000 level (7% for P; 34.6% for X; 14.8 % for J). On the other hand the profiles of G and M both showed a decrease at the UWL level (6.3% and 15% respectively) but an increase in words at the NIOL level (13.2% in the case of G and 6.7 % for M).

Examination of the Beyond 2000 profile analysis of the subjects' pre-course writing on their placement tests shows a closer relationship to their overall writing sub-test results than to their reading sub-test scores (Figure 1). According to the placement test results P had the strongest writing ability, followed by G and M, J and X. By comparison, the LFP analysis of the same piece of writing revealed that most beyond 2000 vocabulary was used by P (19%), and then by J (13%) G (12%), M (11%) and X (9%). In terms of the reading component of the placement test, however, J and X had achieved the highest scores, followed by P, M and G.

Analysis of Subjects' Performances on the MC Cloze

The individual scores of the NNS subjects ranged from 50% to 85.7% (Figure 5). G achieved the highest score, followed by X (61%) and P, M and J, with scores of 50% each. The NS subjects' scores, by comparison, ranged from 100% to 71%, with 85.7% as the mean score, and a standard deviation of 9.27%

NNS Subjects

The NNS subjects' group results on each item in the MC cloze indicate that the collocations tested were only selected by a minority: only one subject chose *stock* (Item 2) and two each selected *vivid* (Item 3), *overwhelming* (Item 6) and *sea* (Item 7) (Figure 6). Results were similar for items which primarily tested stylistic choices: only one subject selected *sterilised* (Item 1), and two each chose *conspicuous* (Item 4) and *lingered* (Item 5). Alternatives chosen for other items were grammatically accurate (no-one chose *stayed* for Item 5) and semantically appropriate (according to the scoring criteria) with the exception of Item 1, where two subjects chose *purified*, and Item 3, where two subjects selected *extreme*. In addition to these last responses, the NNS subjects' response patterns differed most significantly from those of the NS norming group (Figure 2) in their choices of the words *origins* (Item 2), *intense* (Item 3) *noticeable* (Item 5) and *mass* (Item 7).

NS Subjects

In all but one item (Item 5) the L1 subjects' responses validated the researcher's intuitions as to those items that would be considered collocational, and those that would be considered stylistically or semantically appropriate. In the case of collocations (Items 3, 6 and 7), the great majority of L1 subjects recognised 'vivid memory', 'overwhelming anxiety' and 'sea of faces' as collocates. In the case of Item 2, 'of British or Irish *stock*' was accepted as a fixed lexical phrase by a half of the respondents, while 45% or nine respondents chose *ancestry* as being stylistically more refined and semantically more connected with Britain, and one respondent chose *origins*. In terms of items requiring stylistic/semantic choices, most respondents expressed dissatisfaction with all alternatives for Item 1; a number suggested 'sterile' as being more appropriate than *sterilised*, although the majority of subjects still chose the latter alternative. In the case of Item 5, 55% of respondents rejected the author's choice of *lingered* and selected the stylistically more neutral *continued* instead. Respondents argued that *lingered* was less active, less conscious and weaker than *continued*.

Summary of results

In short, results of the LFP analysis using the Beyond 2000 profile indicated that the subjects' beyond 2000 level vocabulary use ranged between 10% and 20% in their pre-test compositions, and between 10% and 30% in their post-test profiles. The subjects' pre-test profiles were consonant with their overall placement writing proficiency scores, but not with their placement reading scores.

The LFP results using the second profiling method showed that increases in the subjects' beyond 2000 vocabulary use in post-test compositions corresponded to increases in vocabulary at the UWL level. Increases in the NIOL level were also registered in the case of the less proficient students, as well as in the post-test compositions of G and M, whose beyond 2000 profiles had in fact declined.

Analysis of the MC Cloze results revealed that the individual NNS subjects' scores were noticeably different from their active vocabulary profiles (as shown by the LFP analyses). While NNS subjects' scores ranged from 50% to 85.7%, they generally found it difficult to successfully make choices testing collocations and stylistic

appropriateness. Consequently, their group response patterns differed considerably from those of the NS subjects.

The baseline data provided by the NS corpus in relation to collocational items appeared to support the validity of the MC Cloze test used. Conversely, items testing stylistic differences or a combination of semantic and stylistic appropriateness proved to be controversial.

Discussion

On the one hand, the results of the LFP analysis add further support to the relationship between writing proficiency and vocabulary claimed by Laufer and Nation (1985) and others (see p.4 above). Analysis of the results according to the second condensed profiling method also suggested that increases in vocabulary at the UWL level was a more reliable indicator of vocabulary development than increases at the NIOL level. On the other hand, given that the strongest increases in beyond 2000 level vocabulary were registered in the results of two of the three subjects who took their placement tests three to four months before the LAAL course, the instrument did not appear to be sufficiently sensitive to detect changes over a relatively short period of time (14 to 16 weeks). An alternative explanation may be that timed 'free' compositions are not conducive to subjects using more sophisticated vocabulary. This was suggested by Laufer (1998: 257) as an explanation of non-significant LFP outcomes in a 1998 study. Melka (1997: 95-96) also discusses the negative impact on vocabulary production of 'avoidance', for either cultural or linguistic reasons (uncertainty about the form, syntax, meaning of words). A further possibility is that individual subjects whose beyond 2000 vocabulary levels appeared to have decreased in the analysis of their post-course compositions may have found the post-course composition topic less inspiring than the pre-course topic on which they wrote. M admitted this to be the case in her post-course interview, and it may also explain G's results. Nevertheless, if a difference in the subjects' responses to the two composition topics was an operative factor, then the claimed topic-independence of the LFP needs to be re-assessed

The MC cloze results, on the other hand, raise a number of issues about the validity of using this form of vocabulary testing. While the task provided some measure of the test takers' collocational

knowledge, it was less useful as a test of stylistic awareness. In the case of Items 1 and 5, the NS subjects' responses were clearly dependent on the individual subject's interpretation of the text and of the author's intentions. For example, the use of *sterilised* (Item 1) seemed to be particularly judgemental according to a number of respondents who were not convinced that this was the author's intention. Some subjects questioned whether the author, with her multicultural and possibly multilingual background, had in fact intended to write 'sterile'. Other subjects felt that the use of register in the text was mixed, making it more difficult to choose between alternatives like, *lingered* and *continued* (Item 5) which were designed to test stylistic awareness, or between alternatives which were in part designed to test sensitivity to style, as in the case of *stock* and *ancestry* (Item 2).

Overall, the NNS response patterns differed significantly from and were more varied than those of the NS norming group. Given that the NNS subjects' performance on the MC Cloze was only measured once, their results would have had no meaning without the accompanying interview data. Considering their proficiency scores and the results of the LFP analyses, the subjects' individual results on the cloze test were different to what might have been expected, particularly in the case of P. However, their retrospective comments gave a more realistic indication than did the cloze results of the differing degrees of each individual's knowledge of particular words, which ranged from no knowledge or recognition only, to an ability to give referential and affective meanings. Likewise the subjects' level of confidence in their knowledge provided an indicator of their overall proficiency. In addition, their explanations provide an insight into the reasoning and the vocabulary strategies they used—like guessing, referring to the co-text and the extra-linguistic context—to make choices. For example, one possible reason why G and X achieved the highest overall results on the cloze test was that unlike the other subjects, they used guessing as a strategy and managed to score two points each time they did so. In other instances, the less proficient subjects achieved higher scores on the cloze test through elimination of other alternatives, or through attributing incorrect meanings or associations to certain alternatives.

Conclusions

The pilot study described in this paper evolved from an attempt to examine vocabulary testing instruments in terms of their ability to measure curriculum-related vocabulary development of L2 university students who had completed a 48 hour EAP course based on a stylistic approach to literature. The two measures selected for investigation, were the Lexical Frequency Profile (Laufer and Nation, 1985), a discrete test which provided information about the subjects' productive vocabulary size and richness; and a more contextualised, multiple choice, rational deletion cloze test which focused on the subjects' receptive knowledge of collocations, semantic and stylistic appropriateness.

The LFP analyses provided a means of comparing the frequency level and richness of the subjects' vocabulary use in pre- and post-course compositions, and also a useful gauge of the level of difficulty that the MC cloze passage was likely to represent for the subjects. The Beyond 2000 profile and the second condensed profile method highlighted differences in the subjects' lexical use over time. However, the results suggested that the instrument was more sensitive to longer term changes in subjects' vocabulary than to changes which might have taken place over the duration of a twelve week course. In addition, despite the authors' claim that the LFP is topic independent, there was an indication, supported by the evidence from the post-course interview with one of the subjects, that the choice of composition topic might affect the LFP results. Similarly, it seems that the time constraints under which the subjects completed their compositions might have depressed the richness of their vocabulary output, a conclusion reached by Laufer (1998) in another study using the LFP. If this instrument is used in future research, a take-home procedure to be completed without the use of a dictionary might produce more realistic results. It would also seem worth using the same composition topic on the pre- and post-tests. A further observation is that future research in this area should compare the performances of NNS subjects with those of NS subjects of the same ages and at the same educational levels, since this would provide a relative meaning to the terms 'intermediate' or 'advanced' L2 learner.

The MC cloze test was the most exploratory aspect of this study. It represents an attempt to devise a contextualised test of collocations and of stylistic awareness. The cloze test itself, although limited in

scale, did not provide valid information about the NNS subjects' lexical knowledge. Given the potential problems it demonstrated, it may be that this type of cloze is much more valuable as an awareness-raising activity and as a springboard for discussion, than in a testing context. On the other hand, the NNS subjects' commentary on the cloze task provided a more accurate indication of the depth of their receptive knowledge of particular words and an insight into the types of strategies they used in making their responses.

Likewise, the NS subjects' performances and commentaries emphasised the problems associated with the use of a clozed text, and particularly a literary cloze. The NS responses demonstrated that the cloze task was more successful in eliciting collocational than stylistic knowledge. Responses to items designed to test stylistic awareness, semantic appropriateness and even collocations depended very much on the reader's interpretation of the text. As Carter says, "What is 'normal' and possible will always be a matter of stylistic choice and relative to a dynamic and negotiable interactive context" (1998: 52). Conversely, it is because literature is especially rich in these "instantial" relationships between words (Carter, 1988: 177) that literary cloze class activities are so useful for sensitising students to the difference between a 'unique' arrangement of words and a more 'normal' use.

The study reported in this paper was small in scale and had a number of limitations: the size of the NNS sample, making generalisation impossible; the reduced number of deleted items on the cloze test; the presence of only one measurement of the subjects' performances on the MC cloze; the need for reference to large corpora to provide more objective norming data on collocations. Nevertheless, the study was informative in drawing attention to the value and the weaknesses of the instruments tested. In this sense, it provides directions for further exploration into ways of testing curriculum-related lexical development in the future. In a small way, the study adds to the growing body of exploratory studies seeking ways of measuring the acquisition of vocabulary knowledge.

References

Astika, G.G. 1993. Analytical assessment of foreign students' writing. *RELC Journal 24*: 61-72.

Bahns, J. & Eldaw, M. 1993. Should we teach EFL students collocations? *System, 21, 1:* 101-104.

Carter, R. 1988. Vocabulary, cloze and discourse. In Carter, R. & McCarthy, M. *Vocabulary and language teaching.* London: Longman: 161-180.

Carter, R. 1998. (2nd ed. 1999*). Vocabulary: applied linguistic perspectives.* London: Routledge.

Channell, J. 1988. Psycholinguistic considerations in the study of L2 vocabulary acquisition. In Carter, R. & McCarthy, M. *Vocabulary and language teaching.* London: Longman: 83-96.

Chapelle, C.A. 1994. Are C-tests valid measures for L2 vocabulary research? *Second Language Research, 10, 2:* 157-187.

Cohen, A.D. 1984. On taking language tests: what the students report. *Language Testing, 1:*70-81.

Engber, C. 1993. The relationship of lexis to quality in L2 compostions. Paper delivered at the 27th TESOL Convention, April 13-17: Atlanta, Georgia.

Isaac, A. 1998. Literature in the language classroom: an Australian story. Paper presented at the American Association of Applied Linguistics Conference. March 14-17: Seattle, Washington.

Joe, A., Nation, I.S.P., & Newton, J. 1998. Sensitive tests. Unpublished manuscript.

Jonz, J. 1990. Another turn in conversation. What does cloze measure? *TESOL Quarterly, 24,1:* 61-83.

Laufer, B. 1998. The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics, 19, 2:* 255-271.

Laufer, B. 1997. What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words. In Schmitt, N. & McCarthy, M. (Eds.) *Vocabulary: description, acquisition, and pedagogy.* Cambridge: Cambridge University Press: 140-155.

Laufer, B. 1995. Beyond 2000: a measure of productive lexicon in a second language. In Eubank, L. Selinker, L. & Sharwood Smith, M. (Eds.) *The current state of interlanguage*. Philadelphia: John Benjamins: 265-272.

Laufer, B. & Paribakht, T.S. 1998. The relationship between passive and active vocabularies: effects of language learning context. *Language Learning, 48*, 3: 365-391

Laufer, B. & Nation, P. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics, 16*, 3: 307-322.

Lewis, M. 1993. *The lexical approach: the state of ELT and a way forward.* Hove: Language Teaching Publications.

Linnarud, M. 1986. *Lexis in composition.* Lund: Lund Studies in English.

Meara, P. & Jones, G. 1990. *Eurocentres Vocabulary Size Tests 10KA.* Zurich: Eurocentres Learning Service.

Meara, P. 1983. Word associations in a foreign language: A report on the Birbeck Vocabulary Project. *Nottingham Linguistic Circular, 11*: 29-38.

Melka, F. 1997. Receptive vs. productive aspects of vocabulary. In Schmitt, N. & McCarthy, M. (Eds.) *Vocabulary: description, acquisition, and pedagogy.* Cambridge: Cambridge University Press: 85-102.

Moon, R. 1997. Vocabulary connections: multi-word items in English. In Schmitt, N. & McCarthy, M. (Eds.) *Vocabulary: description, acquisition, and pedagogy.* Cambridge: Cambridge University Press: 40-63.

Nation, I.S.P. 1990. *Teaching and learning vocabulary.* Boston: Heinle and Heinle.

Nation, I.S.P. & Waring, R. 1997. Vocabulary size, text coverage and word lists. In Schmitt, N. & McCarthy, M. (Eds.) *Vocabulary: description, acquisition, and pedagogy.* Cambridge: Cambridge University Press: 6-19.

Read, J. 1999. A new framework for second language vocabulary assessment. Paper presented at the 12th World Congress of Applied Linguistics, AILA, August 1-6: Tokyo, Japan.

Read, J. 1997. Vocabulary and testing. In Schmitt, N. & McCarthy, M. (Eds.) *Vocabulary: description, acquisition, and pedagogy.* Cambridge: Cambridge University Press: 303-320.

Read, J. 1995. Refining the word associates format as a measure of depth of vocabulary knowledge. *New Zealand Studies in Applied Linguistics, 1:* 1-17.

Read, J. 1993. Development of a new measure of L2 vocabulary knowledge. *Language Testing, 10,* 3:355-371.

Schmitt, N. 1999a. Exploring native and non-native intuitions of word frequency. *Second Language Research, 15,4:* 389-411.

Schmitt, N. 1999b. The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing, 16, 2:* 189-216.

Schmitt, N. 1998a. Measuring collocational knowledge: key issues and an experimental assessment procedure. *ITL Review of Applied Linguistics, 119-120:* 27-47.

Schmitt, N. 1998b. Quantifying word association responses: what is native-like? *System, 26:* 389-401.

Schmitt, N. 1998c. Tracking the incremental acquisition of second language vocabulary: a longitudinal study. *Language Learning, 48, 2:* 281-317.

Schmitt, N. & McCarthy, M. (Eds.) 1997. *Vocabulary: description, acquisition, and pedagogy.* Cambridge: Cambridge University Press.

Schmitt, N. & Meara, P. 1997. Researching vocabulary through a word knowledge framework: word associations and verbal suffixes. *Studies in Second Language Acquisition, 12, 1:* 17-36.

Sökmen, A.J. 1997. Current trends in teaching second language vocabulary. In Schmitt, N. & McCarthy, M. (Eds.) *Vocabulary: description, acquisition, and pedagogy.* Cambridge: Cambridge University Press: 237-257.

Soudek, M. & Soudek, L.I. 1983 Cloze after thirty years: new uses in language teaching. *English Language Teaching Journal 37, 4:* 335-39.

Thorndike, E.L. & Lorge, I. 1944. *The teacher's word book of 30,000 words*: New York: Teachers College, Columbia University.

Viberg, A. 1993. Crosslinguistic perspectives on lexical organization and lexical progression. In Hyltenstam, K. & Viberg, A. (Eds.) *Progression and regression in language: sociocultural, neuropsychological and linguistic perspectives.* Cambridge: Cambridge University Press: 340-385.

Wesche, M. & Paribakht, T.S. 1996. Assessing second language vocabulary knowledge: depth versus breadth. *The Canadian Modern Language Review, 53,* 1:13-40.

Xue, G. & Nation, I.S.P. 1984. A university word list. *Language Learning and Communication, 3:* 215-29.

Appendix 1: MC Cloze Task

More recently, the impact of multiculturalism and overseas travel is beginning to show.

The questions tend to be more like: 'Is your name Thai, Sri Lankan, Indian or Filipino?'

A librarian once said to me, 'Oh, I've been waiting for you to come in, so I could ask you about your name. Are you related to the Indian philosopher, Swami Vivekananda?'

'No, I don't think so. It's quite a common name,' and I add an exaggeration. 'Like Smith.'

As an adolescent growing up, my name caused me great embarrassment and I hated anything that was Asian. How I wished to be Mary Smith. I grew up in the _____[1] environment of Canberra, where the population was white, predominantly of

British or Irish _____[2], with a sprinkling of European migrants. One of my most _____[3] memories from my childhood was the difference in physical appearance between myself and those around me.

When I walked in the streets, I was very conscious of other people's stares. Perhaps it was the unusual sight of my mother dressed in a sari or the novelty of an Indian woman with a Chinese child which made us _____[4].

The effects of this self-conscious childhood _____[5] into my adult life. When I began to give lectures in the course of my work, I experienced an _____[6] sense of anxiety, almost reaching phobic level. I endured this for some time, until I remembered the experience of people staring at me in my childhood. When I stood in front of an audience, and was confronted with a _____[7] of 'staring' faces, I imagined the hostility and condescension I had experienced then. *

* Extract from 'Beyond Limbo', a short story by Kitty Vivekananda. In Mycak, S. & Baker, C. (Eds.) 1997. *Australian mosaic: an anthology of multicultural writing* Melbourne: Heinneman.

Alternatives and scoring schedule

Item	Alternatives	Score	Item	Alternatives	Score
1	purified	0	5	lingered	2
	sterilised	2		continued	2
	refined	1		stayed	0
2	stock	2	6	enormous	1
	ancestry	1		extreme	0
	origins	0		overwhelming	2
3	intense	1	7	sea	2
	vivid	2		crowd	0
	marked	0		mass	1
4	visible	0			
	noticeable	1			
	conspicuous	2			

Appendix 2: Figures

Figure 1   Placement proficiency scores: reading, writing and vocabulary/grammar sub-scores
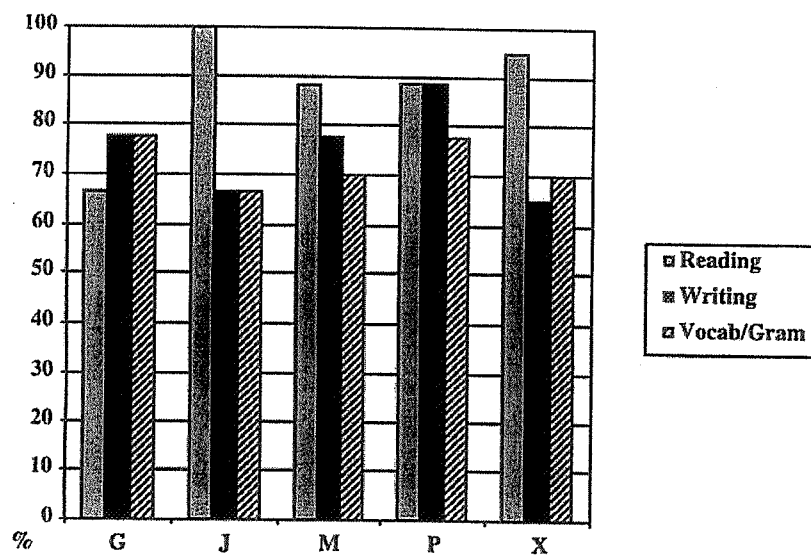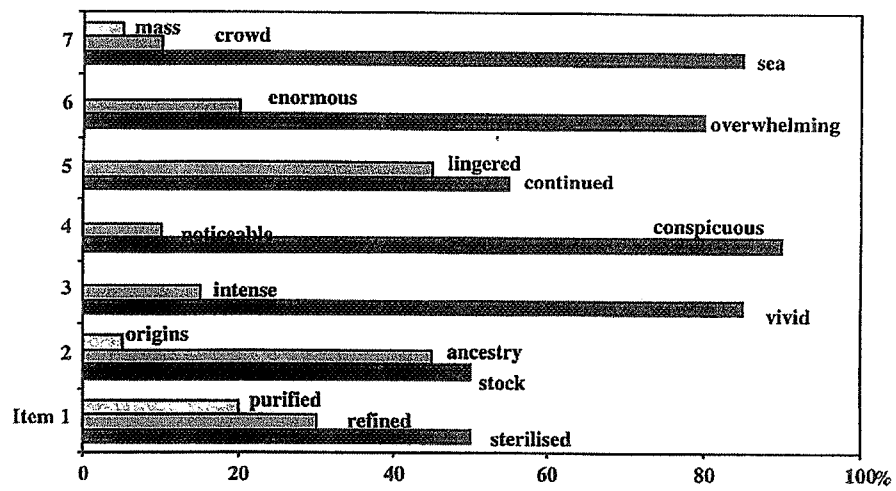
Figure 2   NS responses to MC cloze

Figure 3   Condensed pre- and post-course Lexical Frequency
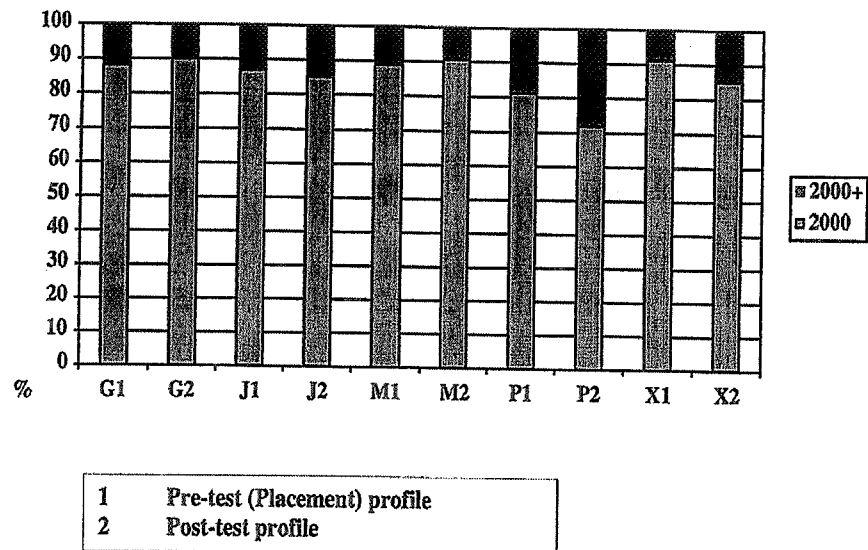   profiles: method 1 (Beyond 2000)

1	Pre-test (Placement) profile
2	Post-test profile

Figure 4   Condensed pre- and post- Lexical Frequency profiles: method 2

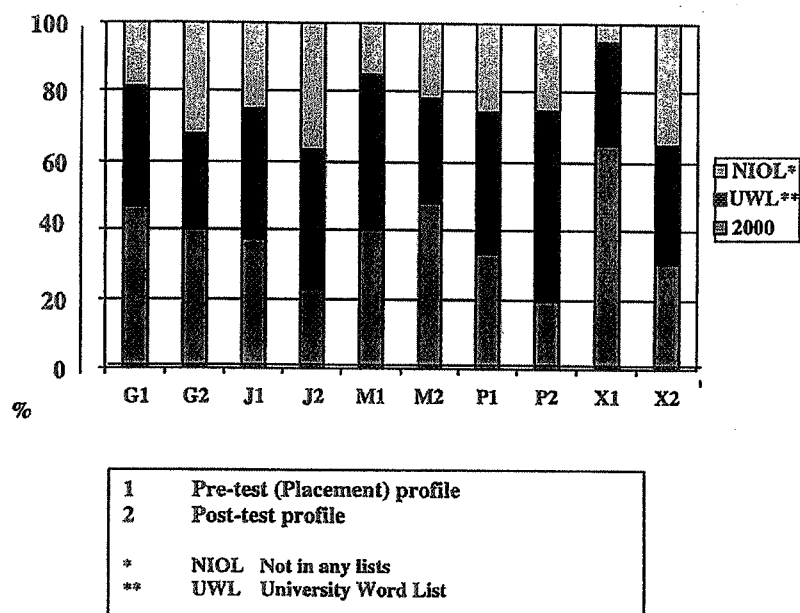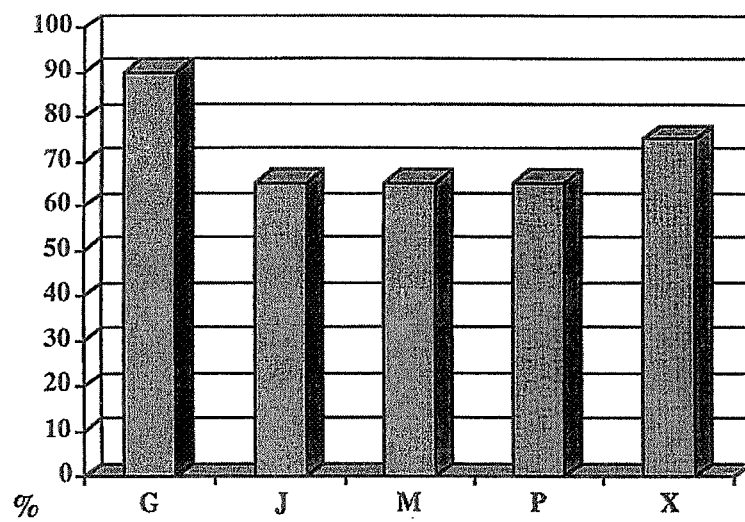| 1	Pre-test (Placement) profile
2	Post-test profile
*	NIOL Not in any lists
**	UWL University Word List

Figure 5   NNS subjects' scores on MC cloze

Figure 6   NNS responses to MC cloze