VIAL

Vigo International Journal
of Applied Linguistics

UNIVERSIDADE
DE VIGO

VIAL. Vigo International Journal of Applied Linguistics.

Este volume foi publicado cunha axuda da Dirección Xeral
de Investigación e Desenvolvemento da Xunta de Galicia

Printed in Spain - Impreso en España
I.S.S.N. 1697-0381
Depósito Legal: VG-935-2003
Imprime e maqueta: Tórculo Artes Gráficas, S.A.

VIAL

Vigo International Journal of Applied Linguistics

Number 1 - 2004

Editors:
Rosa Alonso
Marta Dahlgren

Exploring the validity of a test of productive vocabulary

Tess Fitzpatrick and Paul Meara
University of Wales, Swansea

Abstract

Lex30 is a test of productive vocabulary which uses a word association task to elicit a lexically rich text from the learner. This text is then evaluated according to the number of infrequent words which it contains. Initial studies (Meara and Fitzpatrick 2000) indicated that Lex30 scores might correlate with general L2 proficiency. This paper explores the reliability and validity of the test through a test-retest study and two concurrent validity measures, one using native speaker data and one using a set of collateral tests. These demonstrate that Lex30 produces reliable results and operates with a degree of validity. The results of these studies lead to suggestions as to how the Lex30 test might be improved to further increase its robustness. Lastly, we discuss ways in which the Lex30 studies have shed light on the construct of productive vocabulary and the complex nature of vocabulary knowledge.

Introduction

A few years ago we reported on the design of a new test of L2 productive vocabulary, Lex30 (Meara and Fitzpatrick 2000, Fitzpatrick 2000). The basic premise of this test was that a representative sample of words could be elicited from the productive L2 lexicon, using a word association task. This sample could then be categorised according to word frequency in order to measure the lexical resource of the test-taker. This method has one important advantage over traditional ways of assessing productive vocabulary, in that the "texts" it generates are lexically very dense. Unlike essays, they contain few function words, and a very high proportion of content words. Our preliminary studies yielded some promising, if inconclusive, results, with test scores correlating significantly with another measure of vocabulary size, and we concluded that the test had "considerable potential as a quick and dirty productive test that might be used alongside other tests as part of a vocabulary test battery." (Meara and Fitzpatrick 2000:28). Since the publication of this report, Lex30 has attracted a certain amount of attention from researchers, both in terms of its potential as a practical testing

measure (Baba 2002, Moreno Espinosa and Jiménez Catalán, 2004), and in terms of the contribution it can make to the growing literature concerning the identification and categorisation of vocabulary knowledge (Rimmer, 2000).

Baba in particular drew attention to the fact that the Lex30 studies which had been reported failed to draw any meaningful conclusions about the validity and reliability of the test. We feel that this is a very legitimate and important criticism, and our recent work with Lex30 has included a number of experiments which aim to redress this. It is our intention here, then, to provide a brief description of the Lex30 test and a summary of our 2000 study, and then to report on three experiments which address test reliability, concurrent validity using native speaker norms, and concurrent validity using collateral test measures, respectively. We will also explore the construct validity of Lex30. In conclusion we will discuss a number of issues which have arisen from these experiments, looking both more closely at the design of the test, and more broadly at the validity and usefulness of the concept which it claims to measure.

Lex30: test design

The Lex30 test comprises a word association task, in which subjects are presented with a list of 30 stimulus words in the L2 (English) and are required to produce up to 4 L2 responses, to each of these stimuli. The stimulus words were carefully selected in order to minimise the influence of receptive vocabulary knowledge on the test scores, to differentiate as much as possible between subjects, and to give subjects as much opportunity as possible to produce infrequent vocabulary items. See Meara and Fitzpatrick (2000) for details of the test format and of the formal criteria used for selection of stimulus words.

All of the items produced by a subject in response to the stimulus words form a corpus which is then processed, and a mark is awarded for every infrequent word a subject has produced. In this test, an "infrequent word" is defined as one which falls outside the first 1000 frequency band (Nation 1984). In the pilot study described below, the Lex30 score represented the total number of infrequent words produced, but in all other studies calculation of the score was refined to represent the number of infrequent words produced, as a percentage of the total number of responses given by that subject. This minimised the influence of corpus size on the test score.

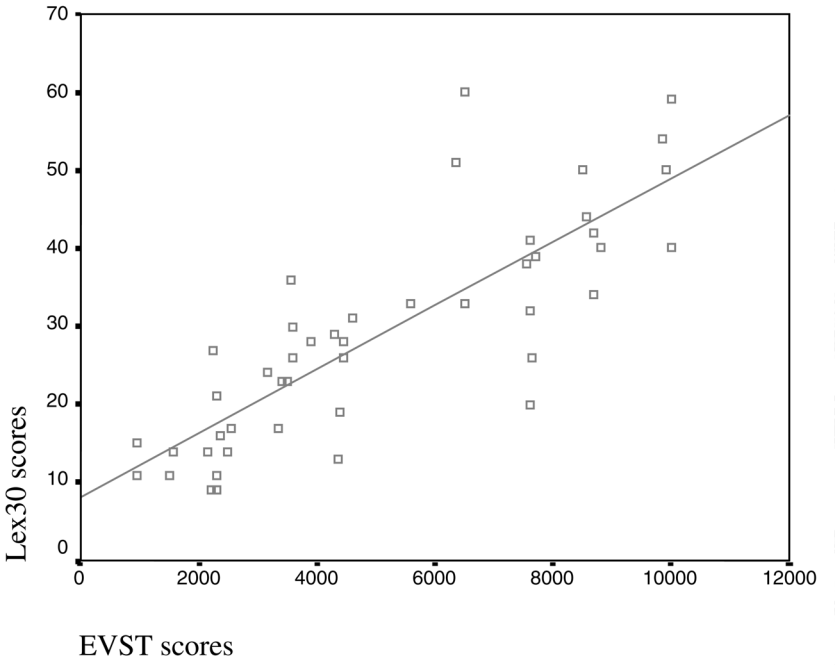
Preliminary validation study

The Lex30 pilot study, which is reported in full in Meara and Fitzpatrick (2000), comprised a comparison of Lex30 scores with scores from another vocabulary test, the Eurocentres Vocabulary Size Test, or EVST (Meara and Jones 1990). The EVST is a test of receptive vocabulary size, and takes the form of lists of words taken at random from appropriate level wordlists. The test is in Yes/No format; subjects are asked to indicate whether or not they "know" the word given. One third of the words given are distractors, i.e. invented words. The subject's approximate vocabulary size can be extrapolated by calculations involving the number of "hits" (correctly recognised words) and "false alarms" (claims to recognise non-existent words). These vocabulary size estimates seem to be a good indicator of overall competence in English as a Foreign Language (Meara and Jones 1988).

EVST was chosen for a control test because it resembled Lex30 in several ways. Both tests involve responses to single word stimuli; for both tests, word frequency is a core concept; both tests take about the same amount of time to administer; and both tests can be easily administered using a computer. A major difference, of course, is that EVST tests word recognition, or passive knowledge, while Lex30 is intended to test productive knowledge. Despite this difference, we felt that the tests were sufficiently similar to be used together in a preliminary investigation into the potential of Lex30; any difference between test constructs could be addressed at a later point, once we had established a general impression of the test's validity.

The subjects used in this pilot study were a group of 46 adult learners of English, from a variety of L1 backgrounds, whose language proficiency ranged from high elementary to advanced level. Subjects completed the Lex30 test and the EVST test within the same week. The Lex30 scores were then compared with the EVST scores of the same subjects; the relationship between these two sets of scores can be seen in figure 1.

Figure 1: Lex30 scores compared with EVST scores



The correlation between these two sets of scores was 0.841 ($p<.01$). This indicates that subjects with a large receptive vocabulary, as indicated by the EVST test, also tended to produce a relatively high number of infrequent words in the Lex30 test, have a relatively large productive vocabulary, and that scores on one of the tests can to some extent be predicted from the other.

These were encouraging results, and indicated that Lex30 had a certain amount of potential as a test tool. However, as was noted in the concluding remarks of the 2000 study, there were still "era number of outstanding issues concerning the reliability and validity of the Lex30 methodology".

Reliability study

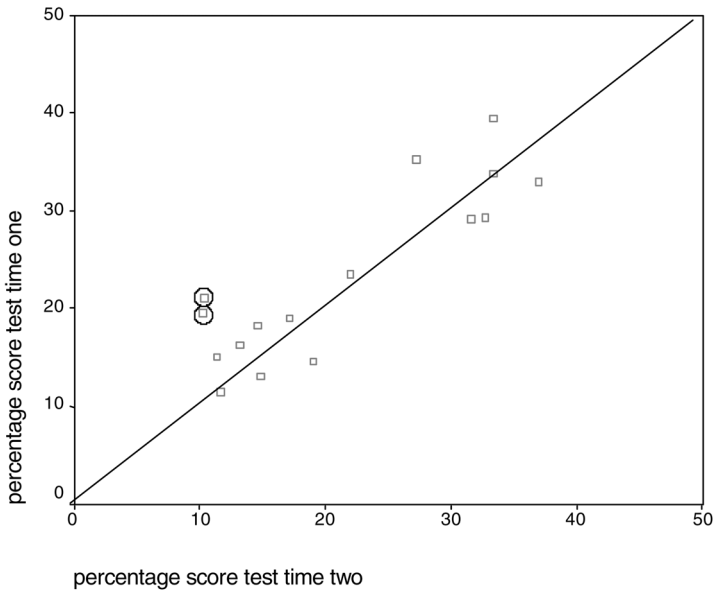
If a test is reliable, it "produces essentially the same results consistently on different occasions when the conditions of the test remain the same" (Madsen 1983:179). A straightforward way to test reliability, then, is to present the same subjects with the Lex30 test on two different occasions, keeping all test conditions consistent, and to evaluate any difference between their scores. One of the crucial features of this test-retest method of reliability assessment is the time

lapse between test time 1 and test time 2. To minimise any "practice effect" (Bachman, 1990), sufficient "forgetting time" must be allowed between test times, but to minimise the effect of improvement (or attrition) in language ability, there should not be too much time between tests. After considering these factors, we decided that a 3-day gap between test times was appropriate.

The subjects used for this experiment were 16 L2 users of English, from a range of L2 backgrounds, and varying in language proficiency from lower intermediate to advanced level. They took the Lex30 test twice, with a 3-day gap between test times. The test and retest scores for each subject are illustrated in Figure 2.

A comparison of means at the two test times gives a t-value of $t=1.58$ ($p=.135$), indicating that there is no significant difference between the two sets of scores. The correlation between the two sets of scores is .866 ($p<.01$). This demonstrates that subjects taking the Lex30 test more than once at a given point in their L2 development will achieve broadly similar scores each time. From this we can propose that the Lex30 test is indeed giving us information about the current state of that subject's lexicon.

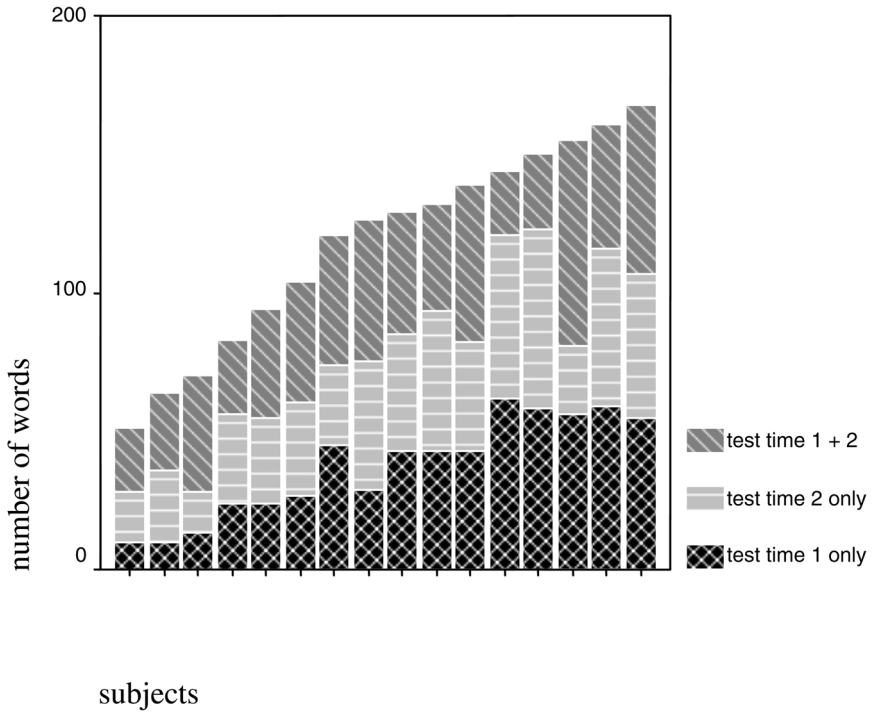
Figure 2: test and retest Lex30 scores for each subject



Clearly the similarity in subjects' scores at test times one and two might be due to them having produced the same words in response to the test task. In order to investigate this, we divided each subject's pair of corpora into three wordlists: words produced at both test times one and two, words produced only at test time one and words produced only at test time two. Figure 3 gives us a visual comparison of illustrates of the relative sizes of these word lists.

Even without examining the statistics for individual cases, we can see that the corpora produced by a subject at test times 1 and 2 were in fact quite different in terms of the actual words they contained. It appears that all subjects demonstrate a tendency to produce new responses on the second test time, regardless of their overall corpus size or their final Lex30 score.

Figure 3: Numbers of words produced by each subject at test time one only, test time two only, and test times one and two



In fact, as the statistics in Table 4 demonstrate, only around half of the words produced at test time one were actually produced again at test time two. However, we know that there is a strong correlation between the Lex30 scores

from the two test times (.866 $p<.01$). These two facts, taken together, allow us to draw an important conclusion about the responses stimulated by Lex30. It appears that, although many of the actual words produced at each test time will be different, the profile of these words will be broadly the same. In other words, subjects are likely to produce similar proportions of infrequent words at each test time.

Table 4: Average number of words produced at test time one, test time two, and test times one and two:

	Test time 1 only	Test times 1 and 2	Test time 2 only
Mean	38	42	39
Sd	18	14	15

This is clearly an important observation. It seems to support the idea that the profiles which result from the Lex30 test task and analysis, are indeed individual to the subject's lexicon; even if the subject produces a different set of response words, the profile remains essentially the same at multiple iterations of the test. This in turn implies that we have succeeded in eliciting a sample of items from the lexicon which are representative, in terms of their inherent frequency, of the overall content of the lexicon.

This experiment, then, has established that the Lex30 test has a high degree of test-retest reliability, and has indicated that the test is successful in eliciting a representative sample of the subject's productive lexicon. However, while "reliability is a requirement for validity" (Bachman 1990 p 238), it is important to recognise that a reliable test is not necessarily a valid test, and we now turn our attention to two experiments which explore the validity of Lex30.

Validity study 1: Native speaker norms

The pilot study described above indicated that subjects perform in a similar way on Lex30 and on the EVST test of vocabulary recognition. However, as we have mentioned, these two tests are based on different constructs – productive vocabulary in the case of Lex30 and receptive vocabulary in the case of EVST, and we should therefore be cautious about validity claims based on this experiment. In Bachman's words, validity is a quest for "agreement between different measures of the same trait" (1990 p240); whether these two tests constitute the "same trait" is arguable.

We are therefore left needing to find other ways of evaluating the validity of Lex30. While one way of doing this is to compare the performance of a subject group on two tests measuring the same trait, a second approach is to look at the performance of two different subject groups on the same test. This approach can assess what Bachman calls "concurrent criterion relatedness", (1990 p 248), with the criterion in question being "level of ability as defined by group membership". Following this approach requires us to identify a group who we know to have a certain level of ability in the trait being measured. Native speakers of English seem to be a sensible choice here; although they disagree as to the actual vocabulary size of a native speaker, researchers agree that the native speaker lexicon will be much larger than that of a non native speaker (Aitchison 1987, Meara 1988, Nation 2001). We can argue therefore that by comparing the performance of a group of native speakers on Lex30 with the performance of a group of non-native speakers, we will be able to evaluate the concurrent validity of the test.

The subjects used for this experiment were 46 adult L1 speakers of English from Britain and North America. The native speaker subjects completed the Lex30 test, and their scores were then compared with those of the 46 non-native speakers' scores which we had obtained from the pilot study (for this experiment all scores represented the number of infrequent words expressed as a percentage of all words produced). The descriptive statistics for native and non-native speaker groups are shown in Table 5.

Table 5: Descriptive statistics for native and non-native speaker Lex30 scores.

	n	Mean	Sd
Native speaker	46	44	7.62
Non native speaker	46	30	9.34

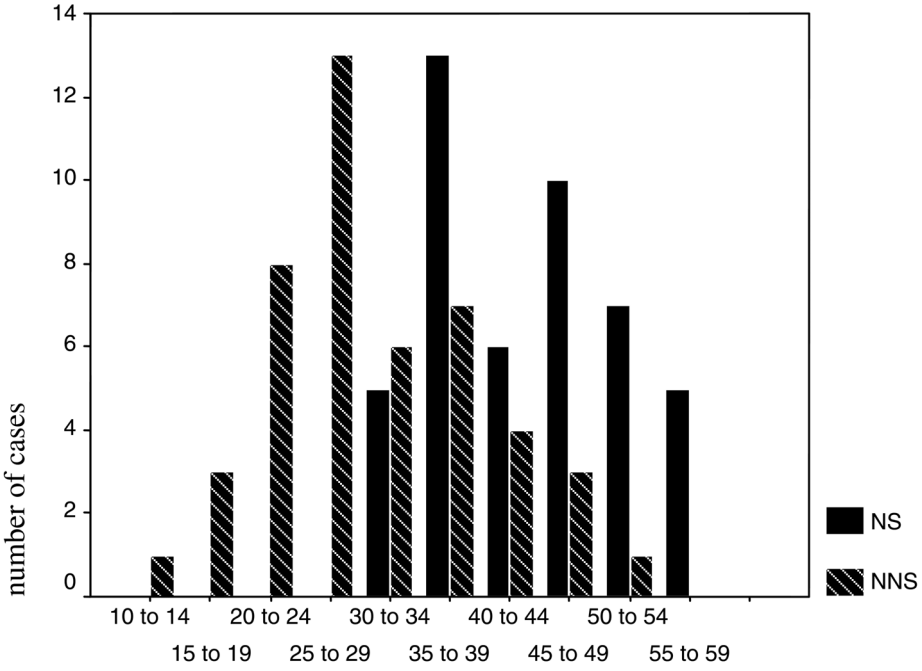
In general, native speakers' Lex30 scores are higher than those of non-native speakers. In fact, the table shows that the mean scores of the two groups differed considerably, with non-native speakers scoring an average of 30 and native speakers averaging 44. An independent samples t-test also indicates that native speakers score consistently higher than non-native speakers taking the test ($t = 7.5$ $p<.0001$).

A closer look at the statistics, though, shows us that the difference between the scores of the two groups is not an absolute one. In fact, as illustrated in Figure 6, which shows the number of native speaker and non-native speaker cases

falling within each band of scores, there seems to be a good deal of overlap between the scores of the groups.

These results raise two important issues about the way Lex30 measures the productive lexicon. Firstly, there appears to be a broad but distinct difference between the scores achieved by native and non-native speakers. Secondly, though, and somewhat contrarily, there is a considerable degree of overlap between the scores of the two groups.

Figure 6: Number of cases falling within score bands

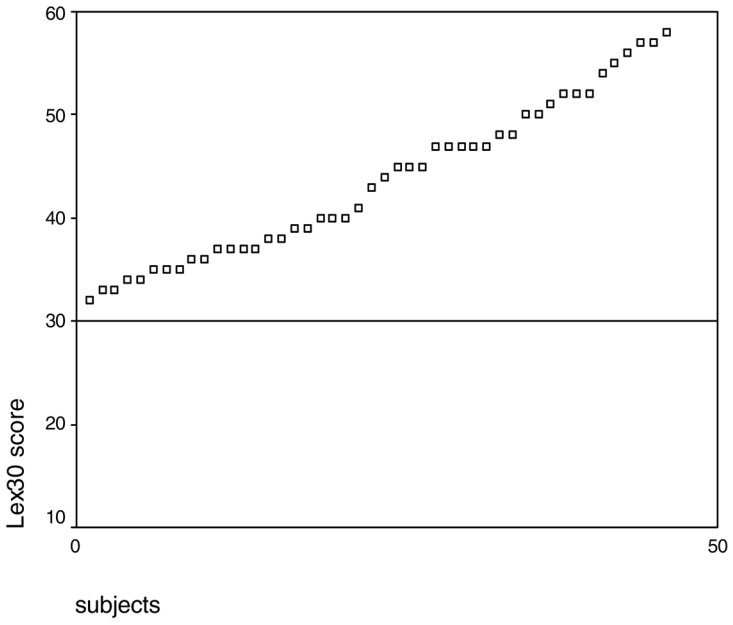


A comparison of the mean scores of the two groups of subjects leaves us in little doubt that native speakers respond to the Lex30 test differently from non-native speakers; they produce a higher percentage of low-frequency words in response to the association prompts. The simple answer to the question of why they do this is, of course, because they can. Our native speakers have a larger lexical resource than our non-native speakers. It is likely that both groups' lexicons will contain most if not all of the 1000 most frequently occurring words in

English, as these are the most commonly encountered and probably the most often used. The lexicons will differ, then, in the number of infrequent words they contain. If we randomly select words from the larger lexicon of the native speaker, we are more likely to retrieve infrequent words than we will from the smaller lexicon of the non-native speaker. In this respect the results of this experiment are very encouraging; we designed the Lex30 test in order to obtain as representative a selection of words as possible from the productive lexicon. It makes sense to assume that the sample from the native speaker lexicon will contain more infrequent words than from the non-native speaker lexicon, and indeed this is the case, indicating that our sampling technique is an effective one.

This conclusion is tempered somewhat, though, by the fact that there is an overlap between the scores of the two groups, with 18 non-native speakers achieving a higher score than some native speakers, and only 6 of the native speaker group scoring higher than the highest scoring non-native speaker.

Figure 7: Distribution of native speaker scores (reference line shows non native speaker average score)



We should perhaps not be surprised about the variation in native speaker scores; while in theory Bachman believes native speakers should provide us with an effective control group, the complexities of their language use can make this

a problematic choice in reality. Bachman warns that: "The language use of native speakers has frequently been suggested as a criterion of absolute language ability, but this is inadequate because native speakers show considerable variation in ability" (1990:39). The abilities which he particularly has in mind are "cohesion, discourse organisation and sociolinguistic appropriateness", and while we had hoped that the discrete and context-free nature of the Lex30 task made it less susceptible to variation, this is perhaps not the case. Despite individual variances, though, the native speaker subjects all scored higher than the average non native speaker score. This is illustrated in Figure 7, where the average non native speaker score of 30 is marked.

Figure 8: Distribution of non native speaker scores
(reference line shows native speaker average score)

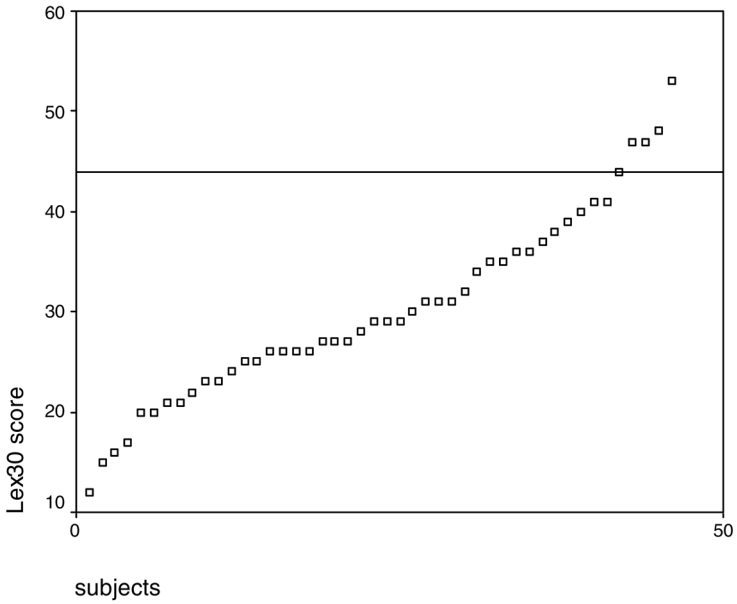


Figure 8 allows us to compare non native speaker subjects' scores with the native speaker mean score of 44. Five non native speaker subjects actually scored higher than the native speaker average. It is helpful to look more closely at those 5 non native speaker subjects with exceptionally high scores. Four of the five subjects are Icelandic secondary school teachers of English, which in itself marks them out as potentially very proficient language users. In our pilot experiment we obtained EVST scores for these subjects, and these are shown in Table 9.

Table 9: Lex30 scores and EVST scores of highest scoring non native speaker subjects.

	Lex30 score	EVST none
nns 1	53	6500
nns 2 (I)	47	9900
nns 3 (I)	47	9850
nns 4 (I)	47	10000
nns 5 (I)	47	7700

(I) = Icelandic teacher of English

These EVST scores are interesting because the EVST test has a ceiling score of 10000; native speakers consistently score between 9500 and 10000. This suggests that, at least for subjects 2, 3 and 4, the Lex30 test, like EVST, is simply not sensitive enough to recognise them as non-native speakers. Subjects 1 and 5 have relatively high EVST scores too, though not in the native speaker range. Subject 5 is a teacher of English, and subject 1 is a very proficient German student who, in terms of tests and coursework, consistently scores higher than his peers in the top level advanced language class. The fact that Lex30 fails to mark these unusually proficient subjects as non-native speakers indicates that it does in fact work well enough to pick out quasi native speakers.

We can conclude, then, that this study demonstrates that the Lex30 test has some validity. Insofar as the design of the test allows, it can distinguish native speakers from non native speakers. For Lex30 to be of practical use, though, it should distinguish between non native speakers of different language proficiency. Our next study addresses this issue.

Validation study 2: collateral tests

A large part of our motivation for devising Lex30 was the dearth of effective tests of productive vocabulary currently available. We have discussed this issue at more length elsewhere (Meara and Fitzpatrick, 2000, Fitzpatrick, 2003). However, we feel that our investigation into the validity of the Lex30 test would be incomplete without making a comparison of its behaviour alongside other tests which claim to measure the same construct, notwithstanding that we have some reservations about the effectiveness of those tests. The final study we will describe here, then, is a test of concurrent validity "examining correlations among various measures of a given ability" (Bachman, 1990, p248). The meas-

ures we have selected for use alongside Lex30 are the Controlled Productive Version of the Levels Test (Laufer and Nation, 1999) and a straightforward translation task from L1 to L2.

The Productive Levels Test, like Lex30, evaluates vocabulary knowledge with reference to word frequency bands. 18 target words are selected from each frequency band, and are embedded in a contextually unambiguous sentence. The first few letters of the target word are given in order to eliminate other conceptually possible answers, and subjects are required to complete the target word. The vocabulary knowledge displayed in the completion of this test is productive in that the subject has to be able to write the word rather than recognise it. It is controlled in that the subject is prompted to produce a predetermined target word, whereas in free productive vocabulary tasks such as composition writing or oral presentation, or indeed Lex30, there is no compulsion to produce specific words. The test incorporates five frequency bands: the 2000, 3000 and 5000 word levels, the University word list level and the 10000 level (Laufer, 1998). Laufer and Nation suggest various methods of scoring the test, but in their 1999 study they calculate scores by counting the number of correct answers given at each level and simply adding them together. This is the method we use here.

The second validation tool used in this study is a straightforward translation task from the subjects' L1, in this case, Chinese. Subjects were given a set of 60 Chinese (Mandarin) words and asked to translate them into English. To minimise the effects of synonyms and homonyms, the first letter of the correct answer for each item was provided. The set of 60 words consisted of 20 randomly selected from Nation's first 1000 frequency list, 20 from the second 1000 and 20 from the third 1000 (Nation, 1984). This meant that the target words were of varying difficulty, and were broadly comparable to the difficulty of the words used in the other tests. The Translation Test is clearly a task of productive vocabulary ability, and unlike the Productive Levels Test has the advantage that it is a context free task, which does not depend on subjects understanding the context the word is provided in. In the scoring of the test, subjects were awarded a point for every target word produced, regardless of the accuracy of spelling.

We selected these two tests as tests of concurrent validity because they share certain characteristics with Lex30:

- all three tests work on the premise that vocabulary can be measured – i.e. that we can, to an extent, quantify the number of words a subject has in their L2 and that this number is somehow meaningful in terms of overall proficiency

- all three are tests of productive rather than receptive vocabulary, requiring subjects to write down words which are prompted in various ways (we should note here that the Productive Levels Test does require subjects to engage receptive skills too, in the comprehension of the context sentence)
- the use of frequency bands is central to the design of all three tests; the Productive Levels Test focuses on subjects' knowledge of words at 5 different word bands, and the translation test on the 1000 – 3000 word bands, and Lex30 awards points for words produced from outside the 1000 level.

55 Chinese learners of English were used as test subjects. The subjects were all undertaking a preparatory "pre-sessional" programme of English language improvement classes in preparation for entry to university in Britain. Their class teachers rated them from intermediate level to advanced, which normally means that we could expect them to know most of the target words in the Translation test and the first two to three levels of the Productive levels test, and all of the cue words in the Lex30 test. The tests were administered during two class sessions, with subjects completing first the Lex30 test task and then the translation task in the first session. In the following day's class, subjects were given the Productive Levels Test.

Table 10: Correlations between test scores

	Productive Levels Test	Translation test
Lex30	0.504 ($p<0.01$)	0.651 ($p<0.01$)
Productive Levels		0.843 ($p<0.01$)

Table 10 shows that there were significant correlations between the results of the three tests. However, the correlations were not as high as we had expected on the basis of the common test factors listed above. While the scores from the Translation test and the Productive Levels test correlate strongly, there is a much more modest correlation between these two tests and Lex30. This suggests that either the tests are in fact measuring different things, or that the tests vary in their degree of accuracy.

Let us first attempt to explain the strong correlation between the Productive Levels test and the Translation test. All the words used in the translation test were from the 3000 most frequent English words. Although the Productive

Levels test targets words from each of 5 frequency bands, in reality the subjects in this experiment struggled to produce any target words at bands higher than 3000; almost all of the correct answers they produced were at the 2000 and 3000 levels. This means that the Productive Levels Test scores reflected to a very large extent - exclusively in many cases - subjects' knowledge of the first 3000 words. This explains the high correlation with our translation test scores; in effect, the two tests were focussing on the knowledge of the same 3000 words. The Lex30 test, on the other hand, takes into consideration – and awards marks for – any words from outside the first thousand. By requiring subjects to produce words spontaneously rather than prompting them to produce pre-selected target words, the Lex30 test can give credit for knowledge of all infrequent words, no matter which frequency band they are categorised in. We know that the Lex30 scores consist mostly of words from the third thousand and beyond, with some contribution from the second thousand band (Meara and Fitzpatrick, 2000). This means that the Lex30 scores are less dependent only on the subjects' knowledge of words in the first three thousand bands, than are the scores generated by the Productive Levels Test and the Translation Test. Clearly, further analysis of this feature of Lex30 is required.

We still need to explain, though, the lack of a strong correlation between Lex30 and the other two tests, and we suggest that this is due to the fact that the tests are measuring different aspects of vocabulary knowledge. An expectation of high correlations between the tests assumes that all three tests measure productive vocabulary knowledge exclusively and completely. Vocabulary knowledge, though, is a rather more complex concept than this implies. To illustrate this, Table 11 lists Nation's aspects of word knowledge (1990), with an indication of which aspects are measured by each of the three tests in this study.

The table indicates that despite their superficial similarities we might expect correlations between the three tests to be modest – they are in fact measuring different aspects of productive knowledge. The modest yet significant correlation between Lex30 and the Translation Test and Productive Levels Test indicates that the tests are operating in the same broad area of knowledge, but the Lex30 test appears to be tapping into different aspects of productive vocabulary knowledge than the other tests.

Table 11: Aspects of Word Knowledge (from Nation, 1990) tested by the Translation test (T), the Productive version of the Levels Test (P), and Lex30 (L).

ASPECT OF WORD KNOWLEDGE (R=receptive, P=productive)			T	P	L
form: spoken form	R	what does the word sound like?			
	P	how is the word pronounced?			
form: written form	R	what does the word look like?			
	P	how is the word written and spelled?	y	y	y
position: grammatical position	R	in what patterns does the word occur?		y	
	P	in what patterns must we use the word?			
position: collocations	R	what words or types of words can be expected before or after the word?			
	P	what words or types of words must we use with this word?			
function: frequency	R	how common is the word?			
	P	how often should the word be used?			
function: appropriateness	R	where would we expect to meet this word?			
	P	where can this word be used?			
meaning: concept	R	what does the word mean?		y	
	P	what word should be used to express this meaning?		y	y
meaning: associations	R	what other words does this word make us think of?			
	P	what other words could we use instead of this one?			y

Discussion

We began our exploration of the Lex30 test by identifying a need for an effective test of productive vocabulary. The design of the Lex30 test seemed to be an attractively simple way of meeting this need; it elicits vocabulary in an efficient way and processes the resulting corpus according to the sort of word frequency criteria which have been accepted as common currency by many language testers. However, the studies described above have left us with some important residual issues to discuss. The first of these are technical issues relating to the design of the test itself.

In order to operate effectively, the Lex30 test has to achieve two broad objectives. Firstly it has to elicit a representative sample of vocabulary from the

productive lexicon, and secondly it has to evaluate this vocabulary in an effective way. Our studies so far have given two major indications that the Lex30 test achieves the first of these aims satisfactorily. Our test-retest study showed that, although the corpora produced by subjects at test times one and two contain many different lexical items, the frequency profile of these corpora are broadly the same. Secondly, our native speaker subjects score higher on average than all but the most proficient non native speaker subjects. These results also indicate that the elicited vocabulary is being measured with some accuracy, too. However, we believe that using a more up-to-date set of frequency bands might improve the accuracy of the Lex30 measure. To this end we are currently engaged in producing a revised version of the test, which uses the JACET 8000 wordlists (JACET 2003).

One of the major advantages of Lex30 as a test of vocabulary is that it is easy to administer. This is especially the case since we have succeeded in automating the data collection stage in the testing process. We have recently produced a programme which automates the processing and scoring of the test (Meara and Fitzpatrick 2004), and will report on experimental studies arising from this new format in due course. Copies of the current version are available from http://www.swan.ac.uk/cals/calsres .

The second major issue to emerge from these studies is a more complex one, and relates to the construct on which the test is based. It seems straightforward to describe the vocabulary which is tested by Lex30 as "productive vocabulary". However, subjects' knowledge of the words they produce could vary widely. For some response words, subjects might only have a threshold level of knowledge, where they know the form of the word and can reproduce it reasonably accurately. For other words, they may have a much deeper knowledge, where they know about its form, use, register, collocations, meaning, associations and so on. This variation is not something that we would expect to find in the data produced by the Productive Levels Test, for example. That test seems to demand knowledge of form, meaning and collocation of target words, as well as understanding of the contextual cue sentence. As Table 11 above exemplified, the concept which we are in the habit of referring to as productive word knowledge, actually encompasses many subcategories of word knowledge, each of which learners will have acquired at varying depths, all of which are interrelated and all of which are in a state of potential change. This is a much more complex situation than, for example, the Free Productive, Controlled Productive and Receptive word knowledge distinction proposed by Laufer. In the light of this complexity it is vital to recognise that these so-called productive vocabulary tests address different aspects of word knowledge.

If it is an overgeneralization, then, to call Lex30 a test of productive vocabulary knowledge, what exactly does it test? Producing a word in response to the Lex30 task certainly implies a minimal level of productive knowledge. In this context a subject does not have to demonstrate any collocational knowledge (the word does not have to be placed in a sentence) or even any semantic knowledge (he is not asked to explain his association link), but some knowledge of form is clearly necessary. Read (2000) distinguishes between two kinds of productive vocabulary knowledge; recall and use. His definitions make it clear that the Lex30 test evaluates recall ability rather than use ability. Recall, he says, is tested when subjects "are provided with some stimulus designed to elicit the target word from their memory", whereas "use means that the word occurs in their own speech or writing" (p.156). This presents us with the problem that use presupposes recall but recall does not presuppose use: we know that a word produced in response to the Lex30 task is known in a "recall" sense, but we have no indication of whether or not a subject can also "use" it.

When we examine the constructs of tests which claim to measure productive vocabulary, then, we find that many of them do not measure the same things at all; productive vocabulary is a misleadingly simplistic label for an extremely complex construct. It seems likely that much more work is needed if we are to develop meaningful tools in this area. In the meantime, though, there is clearly a need, among teachers, learners and researchers, for an effective battery of test tools which can be used to gain an insight into the lexicons of individuals as well as shedding some light on the general behaviour of the L2 lexicon. We feel that the studies we have described in this paper indicate that Lex30 is a robust enough measuring tool to fill an important gap in the battery of tests currently available.

References

Aitchison, J. 1987. *Words in the Mind*. Oxford: Blackwell.

Baba, K. 2002. "Test Review: Lex30" *Language Testing Update* 32: 68-71

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Fitzptrick, T. 2000. "Using Word Association Techniques to Measure Productive Vocabulary in a Second Language" *Language Testing Update* 27:64-69

Fitzpatrick, T. 2003. "Eliciting and Measuring Productive Vocabulary using Word Association Techniques and Frequency Bands". Unpublished PhD Thesis, University of Wales, Swansea.

JACET Basic Words Revision Committee (ed.) 2003. "JACET List of 8000 Words

Laufer, B. 1998. The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics* 19: 255-271.

Laufer, B. and Nation P. 1999. "A vocabulary-size test of controlled productive ability". *Language Testing* 16: 33-51.

Madsen, H.S. 1983. *Techniques in Testing.* Oxford: Oxford University Press.

Meara, P. 1988. "Learning Words in an L1 and an L2". *Polyglot* 9:3, D1-E14

Meara, P. and Fitzpatrick, T. 2000. "Lex30: an improved method of assessing productive vocabulary in an L2" *System* 28: 19-30

Meara, P. and Fitzpatrick, T. 2004. *Lex30 v 2.0*. Swansea: Lognostics

Meara, P. and Jones G. 1988. "Vocabulary Size as a Placement Indicator" In: Grunwell, P. (Ed.), *Applied Linguistics in Society*. CILT, London. pp. 80-87.

Meara, P. and G. Jones. 1990. *Eurocentres Vocabulary Size Test 10Ka*. Zurich: Eurocentres

Moreno Espinosa, S. and Jiménez Catalán, R. 2004. "Assessing L2 young learners' vocabulary: which test should researchers choose?". Paper delivered at *BAAL/CUP Workshop: Vocabulary Knowledge and Use: measurements and applications*. UWE, Bristol.

Nation, I.S.P. (Ed) 1984. "Vocabulary Lists: words, affixes and stems" *English Language Institute Victoria University of Wellington Occasional Paper*, 12.

Nation, I.S.P. 1990. *Teaching and Learning Vocabulary*. Boston: Heinle and Heinle.

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Read, J. 2000. *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Rimmer, W. 2000. "What is it to Test a Word?" *Language Testing Update* 28: 25-27

CALL FOR PAPERS
Deadline for Vial 2, 2005: 1 December 2004
PUBLISHER: Servicio de Publicacións da Universidade de Vigo
EDITORS: Rosa Alonso and Marta Dahlgren (Universidade de Vigo)

NATURE OF THE ARTICLES
Computational Linguistics
Foreign Language Teaching and Learning
Language for Specific Purposes
Language Planning
Second Language Acquisition
Speech Pathologies
Translation

FORMAT OF THE ARTICLES
1. Contributions should be written in English using the software package Word. Three printouts of the article and a diskette should be provided. Title of the paper and name, address, telephone number and e-mail address of the author should be included on a separate sheet. (Submissions by e-mail attachment are also accepted)
2. Articles are not to exceed 25 double-spaced pages (12 pt Times New Roman) including an abstract of 10 lines at the beginning and references. **Please do not include notes.**
3. References should be given in the following format:
   Blakemore, D. 1987 *Semantic constraints on Relevance.* Oxford: Blackwell
   Richards, C. 1985 "Inferential pragmatics and the literary text" *Journal of Pragmatics* 9:261-285
4. All correspondence should be addressed to:

Rosa Alonso or Marta Dahlgren
iialonso@usc.es **dahlgren@uvigo.es**
Universidade de Vigo
Facultade de Filoloxía e Traducción
Lagoas-Marcosende
36200 Vigo Spain