_lognostics
tools for vocabulary research

The complexities of simple vocabulary tests

Paul Meara ***Swansea University***

1. Introduction

Over the last six years or so, I have been working with a number of colleagues on a set of standardised vocabulary tests. My initial motivation for working on this topic was a practical one: I had been doing a lot of research on lexical skills in foreign language learners, but I had no simple way of measuring how well developed the subjects' lexicons might be. What I really needed was a simple measure of vocabulary size, which I could use as a way of loosely controlling the types of subjects that I was working with. Ideally, what we were looking for was a very simple test that could be easily computerised, and would take only a few minutes to run and score. My students and I experimented with a number of simple tests which we thought might serve our purpose, but none of them prove to be entirely satisfactory.

One method which we studied very carefully was a type of test which has come to be known as a **spew test**. In this test, the testees are just asked to produce any words they can think of beginning with a particular letter -- for instance, you might ask someone to produce all the words they can think of beginning with the letter B. Usually the test is limited to two minutes or so. At the end of this time, you count up the number of words that each testee has produced. On the face of it, this test ought to be a fairly good test of how many words people know. Obviously, a person with a very small vocabulary will find this task fairly difficult, while someone with a bigger vocabulary ought to find it rather easy. The obvious inference is that is the more words the testees can produce in the given time, then the more words they know. One obvious advantage of a test like this is that it is not very tightly constrained. Any appropriate response is accepted, and the test does not obviously penalise people who have specialised vocabularies. Unlike most vocabulary tests, therefore, the spew test makes a very few assumptions about the kind of words people ought to know. At the same time, it is a task that a native speaker can sensibly be asked to perform, and in theory this makes it possible to make direct comparisons between native and non-native speakers. Despite these promising characteristics, our spew tests were not a great success. They were taken up by one or two other people (notably Palmberg 1987), who claimed that they were moderately successful, but we always found very low correlations reaching the spew test totals and other measures of vocabulary size (English 1985). We also found that these apparently simple tests raised some serious practical and methodological problems, which we were unable to solve. The main problem was that the number of words a testee produced in a test was clearly affected by individual differences. Some testees were competitive and tried very hard to win; others made less effort and produced lower scores. We might have beeng able to solve this problem by testing the testees in their L1 as well as in their L2, but this merely exacerbated other problems. For instance, the fact the different letters of the alphabet don't all have the same number of words makes it difficult to standardise the test procedure for English; standardising it over a range of languages would have been very much more difficult. And then there was the problem of how to score the replies: should BE count the same as a BICYCLE and BITUMEN? What about BECOME, BECOMES, BECOMING and BECAME? In languages which make extensive use of derivational morphology, questions like this rapidly become a major problem. Furthermore, some speakers that we tested appear not to be familiar with a word games of this sort, and claimed that in their cultures only children took part in these tiresome pastimes: these people were obviously at a serious disadvantage in a timed test. More predictably, some speakers confused Ps and Bs, so that a large number of their responses were just wrong. And so on.

We eventually abandoned this line of approach, and looked instead at ways of measuring lexical skills that were more obviously psycholinguistic in nature. We spent a great deal of time looking at word recognition speed. Our first idea was that it might be possible to produce a standard set of words, say, a hundred or so, reflecting a range of frequencies, and find out exactly how native speakers reacted to these words in a recognition task. The most obvious measure would have been how long it took people to decide that a particular set of letters was a word they knew or not: a lexical decision task. This is a task that has been extensively studied by psychologists, and we know most of the factors that affect word recognition speed in L2, as well as in an L1 (cf. de Groot and Barry 1992). We thought, then, but it would be a relatively straightforward task to collect a "standard" set of words, and assess our learners by looking at how far they deviated from native speaker recognition norms. This idea too fail to work. The main reason why it didn't work was that it needed very sophisticated measuring equipment to detect the very small differences between native speakers and learners. Typically, reaction times to words are in the order of 500 milliseconds -- half a second. In laboratories, it is possible to show that different types of word elicit faster or slower reaction times, but the differences are often as little as 20 or 30 milliseconds -- far too small to show up on the type of equipment that can be used in classrooms. More importantly, however, we found that reaction time to words was not actually a good measure of proficiency: we consistently found, for instance, that English children learning Spanish very often recognised Spanish words considerably faster than native speakers of Spanish did.

Before we abandoned word recognition tasks, we developed a very neat test,whose essential components can be seen in Table 1. In this table, we have two set of items. Each item contains a hidden word. In the first set, the hidden words are all English words; in the second set, the hidden words are all French words. Our impression was that this task was very easy for native speakers -- the words simply stood out from the letters surrounding them. For non-native speakers, the task was very much harder. In our test, the testees saw a set of items like this, displayed on a computer screen, and their task was to press a button as soon as they could say what the hidden word was. This task produced reaction times of about one to two seconds, easily measured by an ordinary home computer, and our initial trials showed that the test discriminated well between native and non-native speakers. Unfortunately, however, our initial tests all used hidden words of six letters, like the ones in Table 1.

Table 1: stimuli for a reaction time test

VAMEGDPMCHOOSEFCDLGP
PESDTMMACCEPTERKDELM
DLGPPURELYLMESSDMLEW
PMGFCDLGPERERGERKDEL
MGFSOURCEPESDTMMERKD
AMEGPDMGFPHRASEPESDT
CDLGPEREJECTELMDKRES
MEHEIGHTGPTDMCAELDON
PGDLAREALLYDREKLRELM
MGPGEFCDREVOLTLEDST
FMPGDLPFDESIGNREMOND

AENPRLBPROPREMLRPITE
ITEPERDREULOALHGISTT
LHCITRISTEISTTENOTLS
TNHXGIESPACEMYAENTPR
PRLMBLRPNEAREMUERHSR
CHELAOLCHEVETLOCHACIS
ENQTLSAGOTPAROLEYMIG
TEULOMESUREALHISTCEN
SLTQNRISQUEEULTIPRLM
AENPRLBPROCHEULOARPM
ETIPRLMBALAVANCENPRL

It proved much more difficult to extend this task to a more normal word list which included words of different lengths. Short words were very hard to identify, while long words were very easy indeed -- the exact opposite of what happens in "normal" word recognition tasks. Our experiments with very long strings of letters containing hidden words were not a success. The reader's reactions to a string like:

pretighunduethobacontmonouvirthrtyfoskadifomaclidopaft

will explain why! Even in the short strings, words beginning with vowels were particularly unsalient. We also had some problems with the letter sequences surrounding the hidden words. These were generated by computer program and occasionally at random collection of letters would appear that made up a word different from the intended target. This sometimes produced two or more English words in the string on the computer screen, and was very confusing for the testees. Eventually, then, we abandoned this line of approach as well.

With hindsight, we were perhaps rather too ready to decide that these lines of inquiry were dead ends. I suspect for instance that the real problem with the spew tests was that we simply counted the number of words the testees produced. This was rather a simplistic thing to do, and we could have used a much more sophisticated mathematical model than this. I now think that if we had looked instead at the rate of production of words, and in particular, if we had looked at the way this rate changes, then we might have been able to develop a very sensitive and powerful test. Similarly, with the hidden words test, I now think that we could have developed a much more sophisticated moving display that exploited the possibilities of the computer screen. However, further developments of this sort outside the scope of this paper. Instead, I will describe yet another test that we experimented with, the test which was much more successful than the abortive efforts I have described so far.

2. The YES/NO test

Our most successful test was, if anything, even simpler than the ones I have described so far. It consists of a set of items, some of which are real words, while the others are imaginary nonexistent words. Some examples are provided in Table 2 work on the next page.

In this test, the testee's task is simply to mark which other words they know. We don't specify exactly what this means, other than to stress that if the testees ensure then the answer should be NO. The test produces two sets of scores: **hits** -- real words that the testee recognises -- and **false alarms** -- imaginary words that the testee claims to know. These two scores allow us to calculate how well the testee can discriminate between the two types of stimuli. We can also use these two scores to produce an estimate of the **true hit rate** -- a figure which takes into account how much testees are guessing, and how far they are prepared to take chances when they think they know a word but aren't sure. In most of our YES/NO tests, the real words population is a random sample from some defined word list, and this allows us to use the true hit rate as an index of how well the testee knows the whole word list. If you score 75 percent on a set of sample tests based on the first 3000 words of the Thorndike and Lorge (1944) list, for instance, then we infer that you know approximately 75 percent of the 3000 words.

We have now used these tests are very extensively in a number of relatively large-scale studies. We have developed several versions of the test, all of which use the same basic methodology. Meara & Jones (1990) is a fully computerised version of the test which produces a single overall vocabulary total. This test takes about 10 minutes to run, during which time it will have tested about 300 words sampling a vocabulary of up to 10,000 items. The 300 word test represents a substantial proportion of a testee's vocabulary, especially if the testee's vocabulary size is fairly small anyway. Results from this

Meara 1994

Yes/No Vocabulary Tests for English
Level 1 test 01

write your name here

what you have to do:

Read through the list of words carefully. For each word:

if you know what it means, write Y (for YES) in the answer box

if you don't know what it means, or if you aren't sure, leave the answer box blank.

1: lack □
2: all □
3: nonagrate □
4: carefully □
5: source □
6: job □
7: least □
8: foreign □
9: into □
10: balfour □
11: business □
12: during □
13: lannery □
14: protect □
15: put □
16: mind □
17: painting □
18: beat □
19: company □
20: oxylate □
21: degate □
22: order □
23: usually □
24: gummer □
25: easily □
26: fall □
27: cantileen □
28: well □
29: in □
30: tooley □
31: ralling □
32: sure □
33: tree □
34: just □
35: happen □
36: contortal □
37: lapidoscope □
38: above □
39: far □
40: glandle □
41: exist □
42: channing □
43: dowrick □
44: mundy □
45: quite □
46: member □
47: part □
48: dogmatile □
49: heart □
50: troake □
51: conversation □
52: project □
53: lauder □
54: aistrope □
55: test □
56: not □
57: interest □
58: could □
59: live □
60: retrogradient □

H: **f:** **Dm:**

test show that it has good test-retest reliability, and the scores we get from it correlate moderately well with a range of other language skills, notably reading comprehension and listening comprehension. As well as this computerised test, we also developed a set of pencil and paper tests that use the same technique. Meara (1992) is a collection of English-language tests which can be used to construct a vocabulary profile of an individual learner. We also developed similar tests for French (Meara 1992b), Spanish (Meara 1992c) and Welsh (Awbery and Meara 1992). The example test in Table 2 comes from Meara 1992.

As an example multiple meanings, take a word like **bank**. This form has at least three separate meanings:
bank -- a place we still your money
bank -- the edge of a river
bank -- what an aeroplane does when it turns.
We also have
to **bank** on somebody -- to rely on them.

A testee who comes across **bank** in a YES/NO test might know all of these meanings, or only one of them. My guess is is that most learners would know **money~bank**, but only a few would know **river~bank**. **Fly~bank** is a meaning that even a lot of native speakers would not know. I have listed these meanings in frequency order, but is worth noting that this frequency ordering is not essential: we could easily imagine a specialist course in English for pilots, where **fly~bank** would be a core item, and **river~bank** might not figure at all. The problem for the YES/NO test is that the less frequent items are difficult to test. We are currently working on a more sophisticated version of the YES/NO test that answers this objection by using pairs of words rather than single items. In this version, some of the pairs have associated links (e.g. tree -- leaf) while others are not linked in this way (e.g. tree -- and grill). The testees' task is to mark the associated pairs. This format allows us to test quite unusual meanings of simple words, as in:

tree -- family
tree -- roof
tree -- shoe
tree -- fell
tree -- money
tree -- life

All of which would probably be recognised as good associations by a highly literate native speaker. Nonetheless, the criticism of our original YES/NO tests remains a good one.

The second objection to the YES/NO tests also needs to be taken seriously. In their present form, it is certainly true that testees can score for words they recognise even when they are not able to use them accurately or correctly. Our only serious defence against this charge is that the tests do appear to correlate fairly well with grammatical accuracy tests. Generally speaking, testees who score very well on the test also have a good grasp of basic syntax and morphology. They may not know for certain how to use all the words in the test, but they certainly know how to use the simpler vocabulary. Nonetheless, this problem too, remains one which we have not yet been able to solve.

I had argued elsewhere (Meara 1990) that the YES/NO tests measure a very basic level of vocabulary skill, and that in some ways this might be an advantage, rather than a disadvantage for the tests. There is, as yet, no agreement about how we can measure word knowledge. A number of competing scales have been put forward for consideration, and in fact a lot has been written about the advantages and disadvantages of each. None of these scales has been turned into a fully workable classification scheme, let alone a workable test: the categories they use do not always apply to all words, and they are capable of being interpreted in different ways, so that it is very difficult to apply them objectively in experimental situations. The YES/NO tests on the other hand avoid this complexity by testing only the most basic of words skills -- the testee's ability to recognise that a particular word form is indeed a word. It is not even necessary for the testee to say what the word means. In some ways, our YES/NO task is even less demanding than "mere passive recognition". What we appear to have identified is **the** basic skill on which all other skills depend. If you cannot even recognise that **tree** is in English word, it is difficult to imagine that you can do anything else with it that might count as vocabulary knowledge. On the other hand, almost any other form of knowledge about a word implies that you are able to recognise it when you meet it, so that test does not make any presuppositions about what kind of knowledge is required for you to know word.

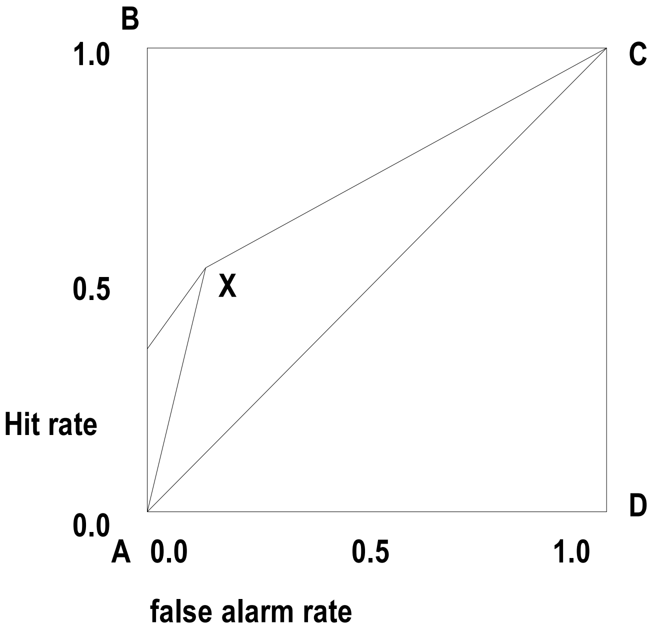
3. Some problems with YES/NO tests.

It seems reasonable to expect that a vocabulary test as simple as the YES/NO test would be a relatively straightforward research tool. We have already seen that the tests make no pretence to be a measure of how well a testee knows a particular word. All we are interested in is the very basic skill of being able to recognise it. Nevertheless, as we have worked with these tests, it has become apparent that they are not nearly so easy decipher as we thought they would be.

Very early on in our work, we developed a set of computer programs that can generate a large number of "equivalent" tests from a basic word list. For instance, if we fed the first thousand words of the Thorndike and Lorge list and a set of matched imaginary words into the programme, it would randomly select a set of matched words and non-words, and compose them into a standard 60 item test. Ideally, tests produced in this way should be equivalent in the sense that a group of testees doing the tests should produce the same mean score and the same standard deviation on both tests. In practice, equivalence has turned out to be elusive. Random pairs of tests tend to correlate moderately well, but in general, we get significantly different mean scores on different tests produced at random. This is not particularly surprising, but with the YES/NO tests these differences seem to persist even when we iron out the major differences by comparing not the scores from a single test but the mean scores of a pair of tests or a set of three or more tests. Even with five randomly selected tests of 60 items, we sometimes find that the means score is significantly different from the mean of a different set of five tests. This slightly worrying. One of the reasons why we get this variation is that the score produced for each test is very sensitive to the number of false alarms testees make, particularly if the testees' hit rate is very low. We are looking at ways of correcting for this sensitivity, but most of the obvious solutions involve increasing the number of imaginary words, and this seems unacceptable on pedagogical grounds: it is rather depressing for testees to have to answer NO all the time. The other conclusion that can be drawn from our problems with equivalence is that the choice of items to go into a test is much more crucial than it is normally assumed to be. A bad choice of words can seriously affect the scoring patterns produced, and even apparently objective selections like the random selections we had been working with, can produce very diverging results.

The second problem that emerged in our tests is that some testees consistently produce negative scores. This statement may need some explanation. The YES/NO tests are scored using a complicated formula. Imagine that we plotted a testee's hit and false alarm rates on a unit square like the one shown in figure 1. In this figure, the point X represents a hit rate of 50% and false alarm rate of 10 percent. The formulae in effect calculates the area of the triangle AXC. The bigger this triangle is, then the better the testee is at distinguishing real words from imaginary words. If the testee is operating at a chance level, then point X will lie very close to the diagonal AC, and the area of AXC see will be very small. The calculation goes on to adjust the actual hit rate to take account of the false alarms, and it does this by projecting a line parallel to AC through X. The point at which this line meets the left-hand edge of the square is the true hit rate.

Figure 1: Scoring a YES/NO test.



It will be immediately obvious that certain combinations of hits and false alarms produce odd results. Any case where the false alarm rate is higher than the hit rate will produce an upside down triangle where the point X lies below the diagonal AC, and in these cases, the true hit rate is less than 0. A surprisingly large number of testees produce results of this kind -- the actual figure varies slightly depending on the proficiency level of the testees. With relatively low-level testees we find between five and 10 percent of people behaving in this way, but occasionally we find a much larger percentage than this. In a recent trial in Canada, for instance, we found an entire class -- some 50 testees -- performing in this way.

It is very difficult to know how results of this sort should be interpreted. The literal interpretation of the data would be that these testees know fewer than zero words, but an interpretation of this sort doesn't make a lot of sense. It is also obvious that these testees do indeed know some words, but the test is failing to pick this knowledge up. What seems to be happening is that the testees are systematically underestimating their knowledge of real words, and overestimating their knowledge of the imaginary words. It is difficult to think of a psychological mechanism that would make sense of this strategy. Our normal practice with data of this sort is to ask the testee to repeat the test, but this time to say YES only in cases of absolute certainty. Even this doesn't entirely eliminate the problem, however. Where a repeat test is not possible, we normally reject the data and eliminate the testee -- an ad hoc solution, and therefore somewhat unsatisfactory, as well as draconian. An alternative solution that we have used once or twice is to weight the false alarms so that a small number of false alarms is penalised relatively lightly, and the full penalty only applies when a significant number of false alarms is produced. Since most testees produce very low numbers of false alarms anyway -the distribution of false alarm responses is normally very close to a Poisson distribution with a mean of 1.5 or so - this solution does not distort the raw data very much, but we have not yet found a weighting procedure that is properly motivated, so that this solution to remains in unsatisfactory one.

A third problem with the YES/NO tests has emerged recently from test on tests we have been running with low-level Arabic learners of English (Al-Hazemi 1993). One of the assumptions underlying the YES/NO test is that the number of false alarms testees make is a reflection of the amount of guessing they have applied to real words. A large number of false alarms suggest that the actual hit rate overestimates the real hit rate by a substantial amount. If this assumption is true, then the number of false alarms that testees produce should correlate with the number of times they say YES to a real item but were actually mistaken. We recently tested this assumption with our Arabic learners and found that it was false. We gave the testees a standard YES/NO test, and then went through the items in the test asking them to provide an Arabic translation for each YES response. A substantial number of these translations turned out to be incorrect, as in the data in Table 3.

Furthermore, when we plotted the number of incorrect responses against the number of false alarms each testee produced, the resulting correlation was far from significant. A large number of translations appeared to have resulted from misreadings of the stimulus words -- a characteristic that we have noted in L1_Arabic speakers in other tests too (Meara and Ryan 1991). And the moment, we don't know whether this problem is one that is specific to very low-level learners, who might be expected to make mistakes of this kind. If this turned out to be the case, then would need to establish why the tests don't work properly at these low levels, and how big a vocabulary the testees need to have before the YES/NO test is a reasonable assessment tool. This would be a significant limitation on the applicapplicability of the tests. On the other hand, it is possible that the problems we have found with these

Table 3: Examples of misinterpreted stimulus words

Target words	**translated into Arabic as:**
chicken	kitchen
careful	clever
finish	fishing
hide	head
pillow	blue
hard	heart
repeat	rabbit
fight	flight
mountain	maintain
tool	tall, tail
sign	sing
storm	steam
basket	biscuit
serious	scissors
invent	infant
dig	dog
pay	buy
cruel	girl, curl, cereal
tidy	tight, today
shake	check
hill	hell, hole
board	bored beard
blow	below, blue

data from Al-Hazemi (1993).

testees are an L1-specific effect, which may not occur with testees from the language groups.

At the moment we are inclined to believe that the YES/NO test is relatively free of L1 effects, unlike most other vocabulary tests, but there are some hints that this might not be the case. We have noted on a number of occasions, for instance, that native French speakers produce odd results without standard YES/NO tests. Generally speaking, the results of our tests correlate fairly well with tests of reading comprehension and listening comprehension, for example. These correlations are modest -- with most L1 groups they work out between .7 and .8 (Meara and Jones 1989). However, we have consistently found that groups of native French speakers produce very much lower correlations that this, usually around .5, and similar results have been found by other people using our tests in Canada (e.g. Wesche and Paribakht 1993). The most likely explanation for this results is that the close relationship between English and French vocabulary makes the test unreliable. The obvious culprit -- the high proportion of cognates in English and French -- does not seem to be the immediate cause of our problem, however (Meara, Lightbown and Halter 1993). Our best guess at the moment is that the problem has something to do with overlaps between orthographic neighbourhoods in the two languages: French words often look like English words and vice versa, even when they are not cognates, and the chances of a French words also being an English word are much higher than would be the case in Spanish or Italian and English. Again, if this conjecture turns out to be well founded, then it will be a further significant limitation on the general applicability of the YES/NO tests. If both Arabic speakers and French speakers behave "peculiarly" on our tests, then it is more than likely that other speakers will too. It obviously does not make sense to calibrate a set of standard tests separately for every possible L1, and the makes even less sense to take into account the effects of other languages that testees might know besides their L1.

4. Discussion

So far, I discussed YES/NO tests in detail, and I have outlined some of the problems that we have identified in trying to use them as measures a vocabulary knowledge in non-native speakers. This is not been merely a ritual breast-beating. I think that the problems I have highlighted are actually much more serious than they appear at first sight.

Over the last few years I have been monitoring very closely the development of empirical research on vocabulary acquisition in foreign languages, and I've been struck by the very large number of people who have developed one off tests for particular projects. It is very difficult to think of any standard tests which has been used in this area, except perhaps the Kent Rosanoff word association list (Kent and Rosanoff 1910), which has in any case been problematical. The nearest thing we have to a standard test at the moment is Nation's University Word List test (Nation 1990), but even this test has been used by only a handful of people and is far from becoming a standard tool.

Very often, the tests used in experimental studies on vocabulary are very crudely reconstructed. Psychologists long ago constructed word lists whose properties are well understood. Norms for imagability, concreteness, orthographic neighbourhoods, and a wide range of other variables known to affect the way native speakers handle words have all been constructed. These norms are all widely used by psychologists in word recognition tasks, for instance, but they don't appear to play any part in the tests that we applied linguists construct. One might be forgiven for thinking that many of the tests used in studies of L2 vocabulary acquisition appear to have been cobbled together in something of a hurry. Often these tests are small, perhaps 20 or 30 items, chosen haphazardly, but used to support sweeping generalisations about the acquisition of vocabulary in much larger numbers. Furthermore, many of these tests involve testing techniques which are much more complex than a look at first sight.

A simple multiple choice test, for example, looks fairly straightforward, but actually involves a large number of different skills, all of which interact in unpredictable ways. Much the same comment applies to gap filling tests, and introspection tests. Furthermore, few of the test used in the recent literature have been standardised in any way. Each team of investigators develops its own tests for its own particular purposes. Some are long, some are short; some include cognates forms, others exclude them; some systematically sample a wide frequency range, others ignore hard or difficult words. Almost always, the scoring is based on very simple mathematical assumptions. Few if any of the tests are reused by other investigators in other environments. Almost always we assume that these home-made tests are valid unreliable instrument of vocabulary knowledge, even when these assumptions are based on fairly flimsy evidence.

In my own tests, I have deliberately tried to reduce these complexities, and I think it is possible to make a strong case for the YES/NO tests as an absolutely minimal vocabulary test, which makes very few assumptions about the nature of vocabulary, and the role it plays in L-2 development. Nevertheless, because we have now used these tests with a wide range of testees, and in a wide range of languages, it is now becoming apparent that even a simple test formats like this one have some surprising complexities. Unexpected L1 effects, surprising individual differences, and more predictable proficiency level effects were all seem to influence the way testees approach the YES/NO test, and clearly influence the way the test works.

If these effects emerge so clearly with our YES/NO tests, which are deliberately kept as simple as possible, then it is more than likely that they must also influence the way other vocabulary tests work too. We have become aware of the problems facing YES/NO tests because we have exposed these tests to a lot of very detailed scrutiny, and their shortcomings are becoming increasingly obvious. This systematic scrutiny is not often applied to vocabulary test used in research, and this must call into question the validity of a great deal of the research published during the recent resurgence of interest in vocabulary acquisition which has relied on one off tests whose characteristics are not well understood.

The obvious solution to this problem seems to be for us to develop a set of standardised vocabulary tests, whose characteristics are well understood, tests that are standardised over a very wide range of languages and learner types. Until we do this, those of us who are interested in measuring the growth and development of vocabulary will be very much like surveyors working with tape measures made of elastic.

References

Al-Hazemi, H
Low-level EFL vocabulary tests for Arabic speakers. Unpublished PhD thesis University of Wales. 1993
Awbery, GM and PM Meara
Graded Welsh vocabulary tests. Swansea. Centre for Applied Language Studies. 1992.
English, F
Measuring vocabulary in non-native speakers in English. Unpublished MA thesis, Birkbeck College, London University. 1985.
Kent, GH and JA Rosanoff
A study of association in insanity. *American Journal of Insanity* 67(1910)37-96, 317-390.
Meara, PM
Some notes on the Eurocentres vocabulary size tests. In: **J Tommola**, (Ed.) *Foreign Language Comprehension and Production.* Turku: AfinLA. 1983. 103-113.

Meara, PM
EFL Vocabulary Tests. Swansea. Centre for Applied Language Studies. 1992.
Meara, PM
Language screening tests. Swansea. Centre for Applied Language Studies. 1992b..
Meara, PM and G Jones
Vocabulary size as a placement indicator. In: **P Grunwell** (Ed.) Applied Linguistics in Society. London: CILT. 1988.
Meara, PM and G Jones
The Eurocentres Vocabulary Size Tests: 10KA. Zurich: Eurocentres. 1990.
Meara, PM, PM Lightbown and R Halter
The effect of cognates on the applicability of YES/NO vocabulary tests. *Canadian Modern Language Review* 50,2(1994), 296-311.
Meara, PM and A Ryan
The case of the invisible vowels: Arabic speakers reading English words. *Reading in a Foreign Language* 7,2(1991), 531-540.
Nation, ISP
Teaching and Learning Vocabulary. New York: Newbury House. 1990.
Palmberg, R
Patterns of vocabulary development in foreign language learners. *Studies in Second Language Acquisition, 9(1987), 201-220.*
Thorndike, E and I Lorge
The Teachers' Word Book of 30,000 Words. New York: Teachers' College, Columbia University. 1944.
Wesche, M and TS Paribakht
Assessing the vocabulary knowledge: depth versus breadth. Paper presented to the 15th Language Testing Colloquium. Arnhem, 1993.

Notes:
This paper first appeared in: **FG Brinkman, JA van der Schee and MC Schouten-van Parreren** (Eds.) *Curriculum research: different disciplines and common goals*. Amsterdam: Vrije Universiteit. 1994. 15-28.