An alternative to multiple choice vocabulary tests

Paul Meara and Barbara Buxton

Language Testing 1987 4: 142
DOI: 10.1177/026553228700400202

An alternative to multiple choice vocabulary tests

Paul Meara *University of London* and **Barbara Buxton** *Cassio College*

This paper reports a preliminary evaluation of the Y/N technique for producing tests of vocabulary knowledge. The results obtained suggest advantages over the more traditional multiple choice format for testing vocabulary.

I Introduction

Not very long ago, it was fashionable for people to talk about vocabulary acquisition in a second language as a neglected aspect of the L2 learning process. Indeed, a whole range of more or less colourful metaphors on this theme have appeared in print in the last five years. Happily, this neglect of vocabulary is no longer the case, and while vocabulary acquisition may not yet have advanced to the status of flavour of the month, we are obviously heading in that general direction. One important side product of this reevaluation of vocabulary acquisition is that it has made us acutely aware of the inadequacy of our current views on how words are stored and acquired in an L2 and of the inadequacy of the tools we traditionally use to measure mastery of words. This paper will be mainly concerned with discussing a new, but very simple way of measuring word knowledge in an L2. This technique actually opens up some important new possibilities in vocabulary testing and these will be discussed in the final section.

II Background

All language tests are to some extent tests of vocabulary: without some knowledge of what the words appearing in a test mean, it is extremely difficult to perform at all, let alone well. In some public examinations, however, vocabulary knowledge is explicitly tested and a student's vocabulary score contributes a substantial portion of the final score he or she is allotted. The Cambridge First Certificate examination is an instance of this approach. This examination includes a section called 'comprehension', which is essentially a vocabulary test. In this exami-

nation and others like it, the vocabulary is assessed by a series of multiple choice questions, in which candidates have to indicate which of a set of possible choices best fits a given context.

In its simplest form, the multiple choice vocabulary test simply presents the student with a single word and a set of meanings and asks the student to decide which of the given meanings is the correct one. An example of this will be found in Table 1 (question 1a). This example clearly tests for knowledge of the word *fade*. The format can produce items of varying difficulty and refinement through careful manipulation of the distractors. Thus, 1a is fairly easy, while question 1b requires a more detailed knowledge of the word field that *fade* belongs to.

Table 1 Examples of multiple choice vocabulary formats

1a	to *fade* means
	a: to make a loud noise
	b: to get something to eat
	c: to paint something a bright colour
	d: to become quieter
1b	when a light *fades* it
	a: gets weaker
	b: gets stronger
	c: flashes intermittently
	d: turns red
1c	The blue curtains began to ____ after they had been hanging in the sun for two months.
	a: fade b: die c: dissolve d: melt

A more sophisticated version of the same basic idea will be found in question 1c (Table 1). In this item, the student is presented with a sentence which contains a missing word and a choice of four words which might fit the gap. Again, the item tests knowledge of *fade*, but it requires the student to know which words collocate with *fade* as well as the general area of meaning to which *fade* belongs.

Tests such as those illustrated in Table 1 have two major drawbacks. First, at the level of item, there are many reasons for suggesting that the tests are not entirely reliable. It is possible, for instance, for a student who knows *fade* to get the answer wrong because he does not understand the other words which supply the context or the definitions, because he is confused by the syntax of the question, or because he knows one meaning of *fade* but not the one required here. This problem gets more acute with items like 1c where the context is more crucial. Furthermore, with items like 1c, it would be possible for a student who did not know *fade* to get the answer right if he knew enough about the other words to exclude them as likely candidates. If the student knows none of the words and guesses blindly, he has a 25% chance of getting the correct answer, but if he can confidently rule out

one or two of the distractor words then the odds on a wild guess being correct shorten to 33% or 50% respectively.

The second major drawback of this type of test works on a different level. It is fairly obvious that we can test knowledge of individual words using items like those in Table 1, but testing individual words is only a means to an end: what we are really looking for is a test that gives us some indication of the overall size of a learner's vocabulary. In other words, when we use test items like those in Table 1, we are making assumptions about what can be inferred from the results of the test. On balance, we assume, a student who gets 20/50 on a test of this sort probably knows fewer words than a student who gets 45/50. This conclusion does not follow automatically, of course, since one could imagine a case of a student with a very large vocabulary who just happened not to know the items appearing in a particular test set. Typically examination boards minimize this possibility by extensive pretesting of item sets, but this process is costly and time-consuming. Even when such pretesting is carried out properly, it still leaves us with one outstanding problem, however. Suppose that you have a learner, whose total vocabulary size is in the region of 1000 words, and this learner is taking a test like the Cambridge First Certificate Examination, where the relevant section of the examination contains 25 items. These 25 items can be seen as a sample of the learner's actual vocabulary of 1000 items. With 1000 words and 25 items, the sampling rate is one in 40. This figure is not too bad, but as the size of the actual vocabulary increases, the sampling rate gets progressively worse. For a vocabulary of 2000 words, a 25-item test samples only one word in 80; for 4000 words, the sampling rate falls to one word in 160; while for a fairly advanced student with a vocabulary in the 10,000 range, a 25-item test samples only one word in 400. In theory, this problem could be overcome by increasing the number of items in the test, but a vocabulary of 10,000 words would require 250 items to achieve a sampling rate of one word in 40 and this is clearly not practical. In short, it looks as if multiple choice tests of vocabulary might work effectively at lower level of proficiency, but as the learner's vocabulary grows they become increasingly unreliable.

III A simple alternative

The purpose of this paper is to examine an alternative form of vocabulary testing which gets round both sets of difficulties noted above. The idea is not a new one: it was originally developed by Zimmerman, Broder, Shaughnessy and Underwood (1977) for use with L1 speakers, and since then the method has been extensively used by Anderson and Freebody (1983), again with native speakers. To our

knowledge, the idea has not been used before with nonnative speakers, however.

A simple version of the test is shown in Table 2 and the reader is recommended to complete this test before going on.

Table 2

Look through the French words listed below. Cross out any words that you do not know well enough to say what they mean. Keep a record of how long it takes you to do the test.

VIVANT	TROUVER	MAGIR	ROMPTANT
MÉLANGE	LIVRER	IVRE	FOMBE
MOUP	VION	LAGUE	INONDATION
SOUTENIR	SIÈCLE	TORVEAU	PRÊTRE
REPOS	GANAL	HARTON	TOULE
GOÛTER	FOULARD	EXIGER	AVARE
ÉTOULAGE	ÉCARTER	MIGNETTE	JAMBONNANT
DÉMÉNAGER	POIGNÉE	ÉQUIPE	MISSONNEUR
AJURER	BARRON	CLAGE	TOUTEFOIS
LEUSSE	CRUYER	HÉSITER	SURPRENDRE
LAVIRE	SID	ROMAN	CHIC
ORNIR	CÉRISE	PAPIMENT	CONFITURE
GÔTER	PONTE		

In this test a list of French words was presented and the reader was asked to indicate which of them were known. In fact, the list consists of both real French words (which the reader might have known) and imaginary French words (nonexistent words which the reader could not possibly have known). This arrangement allows for four types of response:

	Type of word	
response	real	imaginary
Yes	RY	IY
No	RN	IN

The real words tested were a random sample from the *deuxième degrè* of *Français Fondamental*, i.e., a sample from a vocabulary of about 2000 words. Ideally, a reader who knew all these words would have responded YES to all the real words and NO to all the imaginary word and the cells labelled IY and RN in the diagram would both contain zero. In most practical situations, however, there will be some RN responses and some IY responses – cases where real words are not known and cases where the reader claimed to know an imaginary word. These figures allow a check to be made on how truthful the reader is being. Suppose, for example, that the reader claimed to know all the real words and rejected all the imaginary words: in this case, the fact that none of the imaginary words were accepted means that the

reader probably does know the real words. In the contrary case, if the reader claimed to know all the real words, but also claimed to know all the imaginary words, then it is clear that at best he must be mistaken. Intermediate cases can be sorted out using mathematical models based on *signal detection theory* (Kling and Riggs, 1971). Suppose, for example, that the reader claimed to know 50% of the real words, but also claimed to know 20% of the imaginary words. The IY score would indicate that the reader's score of 50% in the RY cell needed to be adjusted downwards. Signal detection theory models allow us to do this in a principled, non-arbitrary way.

The practical advantages of this Yes/No test (Y/N test) should be immediately obvious. The test is very easy to construct and it requires only a few minutes for the testee to complete it. The test in Table 2 sampled one word in 80 of the target vocabulary (25 real items for 2000 words), but in principle it should be possible to sample learners' vocabulary at much higher rates than this, even when dealing with very large vocabularies. The only remaining question, then, is: does the test work in real life? This question is addressed in the next section.

IV The Y/N test: a preliminary evaluation

In this section we report a study which compares scores on a Y/N test with a more traditional multiple choice vocabulary test of the type used in the Cambridge Proficiency Examinations. This work was undertaken as a preliminary evaluation, aimed at assessing whether the Y/N procedure was deserving of further, more thorough investigation. It is assumed that the multiple-choice test is indeed a good vocabulary test but with the disadvantage that it is difficult and costly to produce. If scores on the multiple choice test correlate closely with scores on the Y/N test, then we can conclude that the Y/N procedure deserves a thorough formal assessment.

Materials

Two tests were used in this study: (a) a multiple choice test of the type used in the Cambridge First Certificate Examination (MC) and (b) a specially constructed Y/N test (YN). The two tests will be found in the Appendix. MC consists of 25 items; YN consists of 100 items, 60 real words and 40 imaginary ones.

Subjects

100 subjects took both tests. The subjects were all following English language courses at Cassio College of Further Education in Watford,

UK. All were over 16 years of age; the mean age of the group was 22 years. The group was predominantly European in origin, but also contained some Arabic and Japanese speakers. All subjects had been in the United Kingdom for at least three months at the time of testing, and were attending classes for five hours a week.

Administration

Both tests were administered on a single day, with a 15-minute break between tests.

Scoring

Two scores were obtained for each subject: an MC score and a YN score. It was not possible to find out exactly how the Cambridge multiple choice tests are marked, since this information is withheld from the public. In the absence of clear guidelines, three points were awarded for each correct answer, and one point deducted for each incorrect answer. This scheme takes account of the possibility that subjects may score correct answers by chance, and effectively penalizes them for guessing.

The score for YN was calculated using a formula provided by Anderson and Freebody (1983):

$$P(k) = \frac{P(h) - P(fa)}{1 - P(fa)}$$

This formula derives directly from stimulus detection theory studies, $P(h)$ (i.e., the probability of making a 'hit') in our study is the proportion of real words that the testee recognises (RY); $P(fa)$ (i.e. the probability of a 'false alarm') is the proportion of imaginary words the testee claims to know (IY). The formula adjusts the RY score downwards if IY is large. $P(k)$ in signal detection theory represents the likelihood of a real target actually being acknowledged: in our study it indicates how many of the target words the testee can be deemed to know.

V Results and discussion

The correlation between YN and MC was satisfactory ($r = .703$, $p < .001$, $n = 100$) given the number of people taking part in the study; the indications are that despite the appa-ent differences between them, MC and YN are measuring largely the same sort of thing.

In a similar study using native-speaking children, Anderson and Freebody (1983) found correlations of .84 between scores on an MC

test and a YN test. Their study thus produced correlations which were rather higher than the figure reported here. The main difference between the study reported here and that of Anderson and Freebody is that the latter were able to use identical items in the two tests. This option was not available in the current study for administrative reasons, but it is obvious that, had it been, the correlations obtained would probably have turned out rather higher.

Furthermore, Anderson and Freebody's study had a homogeneous group of subjects. The subject-group in the current study was not homogeneous. However, the largest single group in the current study was a group of French speakers (n = 18). The correlation between MC and YN for this group was .829, a figure which is very close to that obtained by Anderson and Freebody. This question of homogeneity of the target group is actually more important than it looks at first sight, since it presents some difficulties when it comes to designing imaginary words for the Y/N test. Imaginary words which do not cause any particular problem for, say, a German speaker, might be a source of difficulty for speakers of Romance languages, for instance. Take for example the imaginary word *observement*: it bears no resemblance to any German word, but it does look like a possible word in French or Italian, and so the decision to reject it might be harder for a Romance speaker than it would be for a German speaker. In fact a number of imaginary words used in YN were of this type, and it is possible that this may have made the test more difficult for selected L1 groups. Obviously, it cannot be advisable to use just any old imaginary words in a test of this sort. At the moment, however, it is not clear what criteria, other than orthographic and phonological ones, should govern our choice of imaginary words; further research in this area will be needed if Y/N tests are to be developed seriously.

Table 3 Means, standard deviations and reliabilities (n = 100)

Test	Mean	SD	Reliability (KR 21)
MC	12.57	6.47	.88
YN	26.05	11.93	.91

VI Follow-up

So far we have been discussing the relationship between YN and MC, and arguing that there is a close correlation between scores on these two tests. The next question that arises is: which of the two measures is the better one? It is not possible to answer this question directly with the data available, but one indirect measure was available: 26 of the 100 subjects entered for the Cambridge First Certificate in English

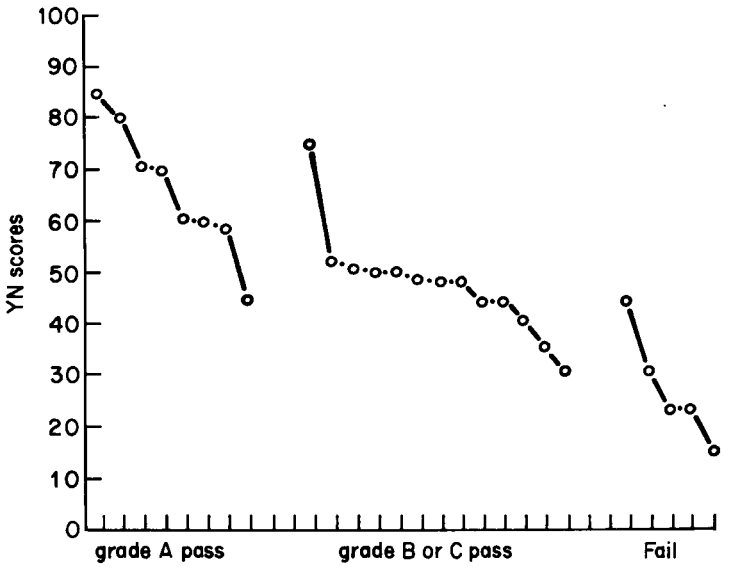

Figure 1 Comparison of First Certificate results with YN scores

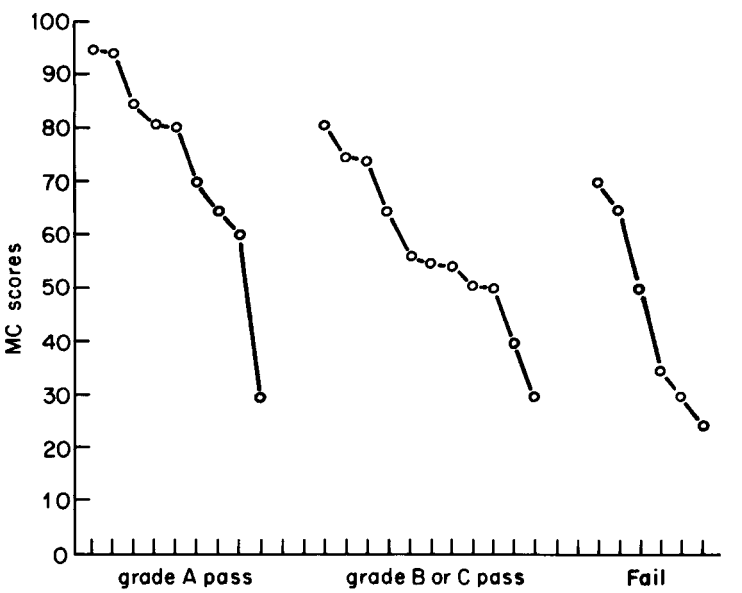

Figure 2 Comparison of First Certificate results with MC scores

examination shortly after taking the two tests reported above. Comparisons between the results of this examination and the two tests are shown in Figures 1 and 2. In this figure, candidates are classified into three categories: grade A Pass, Pass (grades B or C) and Fail. The figure shows clearly that the YN test is much better at discriminating between the candidates than is the MC test, despite the fact that the First Certificate examination actually includes a multiple choice vocabulary test as a major component. Dividing the YN scores into high (61%) +, medium (41–60%) and low (0–40%) categories allowed us to calculate the degree of association between YN and First Certificate scores. This result was significant ($\chi^2 = 13.6, p < .01$ with $2df$). In fact, the YN test correctly predicted the results of all but five of the candidates: two were overestimated and three marginally underestimated. The equivalent figure for MC was not significant.

VII Conclusion

Taken together, the results suggest that the Y/N test may have a lot more going for it than is apparent at first sight. What looks initially like a simple-minded idea turns out, on further examination, to be a remarkably powerful test technique. Given further work on problems like the choice of imaginary words and on variation due to L1 transfer effects, it should be possible to refine the test into a tool which is at least as good as the traditional multiple choice vocabulary test with only a fraction of the developmental costs associated with multiple choice.

However, the real advantages of the Y/N technique do not stop here. A significant advantage of Y/N over multiple choice is that it is possible to use Y/N to test very large numbers of items in a way that is just impossible with multiple choice. This means that it suddenly becomes possible for us to test a significant proportion of the words a learner is expected to know, rather than a tiny sample of them. It takes only a second or so for a testee to decide whether he knows a word or not, and this means that several hundred items can be tested effectively in the space of a few minutes. In itself this is not particularly interesting, though obviously it reduces the risk of penalizing students because of arbitrary sampling. What is important is that the Y/N test format coupled with a formal sampling mechanism makes it possible to come up with a figure which actually quantifies the number of words a student knows. The authors are currently working on a computerized version of a Y/N test which will allow measurement of EFL vocabulary size with a sampling rate of about one word in 10 up to 10,000 words. The prototype version of the test does this in 10 minutes, scores itself automatically and produces very high cor-

relations with an extended test of overall ability in EFL which normally requires an hour and a half to administer. We think this is a significant development, particularly for research purposes, since it will allow researchers to assess the proficiency level of their subjects in a very simple way and help us to get away from the messy and unreliable labels which characterize much of current research.

VII References

Anderson, R.C. and **Freebody, P.** 1983: Reading comprehension and the assessment and acquisition of word knowledge. *Advances in Reading Language Research* 2, 231–56.

Kling, J.W. and **Riggs, L.A.** 1971: *Woodworth and Schlossberg's Experimental Psychology*, Third Edition. London: Methuen and Co.

Zimmerman, J., Broder, P.K., Shaughnessy, J.J. and **Underwood, B.J.** 1977: A recognition test of vocabulary using signal-detection measures and some correlates of word and non word recognition. *Intelligence* 1, 5–13.

IX Appendix

VOCABULARY TEST 1: Choose the word or phrase which best completes each sentence. Put a ring around the correct answer, e.g. (B)

NAME ..... DATE .....
NATIONALITY ..... LANGUAGES SPOKEN .....

1 The blue curtains began to ____ after they had been hanging in the sun for two months.
A fade B die C dissolve D melt

2 Learners of English as a foreign language often fail to ____ between unfamiliar sounds in that language.
A separate B differ C distinguish D solve

3 The wind blew so hard and so strongly that the windows ____ in their frames.
A rattled B slapped C flapped D shocked

4 I have lived near the railway for so long now that I've grown ____ to the noise of the trains.
A accustomed B familiar C unconscious D aware

5 In spite of her protests, her father ____ her train for the race three hours a day.
A let B made C insisted D caused

6 It was impossible for her to tell the truth so she had to ____ a story.
A invent B combine C manage D lie

7 The car had a ____ tyre, so we had to change the wheel.
A broken B cracked C bent D flat

8 She applied for training as a pilot, but they turned her ____ because of her poor eyesight.
A back B up C cover D down

9 The only feature ____ to these two flowers is their preference for sandy soil.
A similar B same C shared D common

10 The play was very long, but there were two ____.
A intervals B rests C interruptions D gaps

11 These old houses are going to be ____ soon.
A laid out B run down C pulled down D knocked out

12 She rang to make an early ____ at the hairdressers.
A order B date C assignment D appointment

13 The law states that heavy goods delivery vehicles may not carry ____ of more than fifteen tons.
A masses B sizes C measures D loads

14 The young soldier ____ a dangerous mission across the desert, although he knew that he might be killed.
A undertook B agreed C promised D entered

15 You must ____ that your safety belt is fastened.
A examine B secure C check D guarantee

16 He ____ a rare disease when he was working in the hospital.
A took B suffered C infected D caught

17 My sister had a baby daughter yesterday, and she is my first ____.
A nephew B cousin C niece D relation

18 When he heard the joke, he burst into loud ____.
A smiles B laughter C amusement D enjoyment

19 The traffic lights ____ to green, and the cars drove on.
A exchanged B turned C removed D shone

20 It is a good idea to be ____ dressed when you go for an interview.
A finely B boldly C smartly D clearly

21 If we go to the market we might find a ____.
A trade B shopping C chance D bargain

22 If he drinks any more beer, I don't think he'll be ____ to play this afternoon.
A skilled B capable C possible D fit

23 That's a nice coat, and the colour ____ you well.
A fits B matches C shows D suits

24 Many accidents in the home could be ____ if householders gave more thought to safety in their houses.
A avoided B excluded C protected D preserved

25 Smoking is a very bad habit, which many people find difficult to ____.
A break B beat C breathe D cough

VOCABULARY TEST 2

NAME ..... DATE .....
NATIONALITY ..... LANGUAGES SPOKEN .....

Tick the words you know the meaning of e.g. milk ✓

gathering	forecast	descent	revenge
strap	conscious	wodesome	heartless
untamed	mudge	topical	mere
loyalment	crope	robber	awkward
flane	possess	weast	loafing
article	amusity	reference	familiar
risent	invaluable	sleme	slight
instructness	heal	guesswork	logless
repeat	influence	jerbal	arrestation
weary	expume	precious	destructive
dismissal	infectory	exclaim	depthen
successment	artificial	judgement	redirect
handle	sloping	rehearsion	sportly
combine	assainful	scatter	basis
strangity	bundle	misled	preferable
magnify	bluck	bathe	unvelop
forgivity	proposal	peculiar	efficiency
arousion	fortake	compire	draven
rejected	flapping	enclose	steady
deformness	inscarce	plode	observement
collar	conversal	allowance	porfume
infect	miggle	eccentric	warness
burdle	whistle	broaden	freath
lodge	turmoil	groppable	henge
recipe	forcement	quietness	levity