A VOCABULARY- BASED GRADED DICTATION TEST

R. L. FOUNTAIN
I. S. P. NATION
Victoria University of Wellington

Abstract

This article describes a dictation test which is graded according to vocabulary frequency. The article describes how the tests are made, used and marked. Four equivalent forms are included in the article. Statistical data on the tests is provided, showing that the tests have high reliability, and are valid tests of vocabulary knowledge. Evidence is provided to show that the four forms are equivalent.

For many years, researchers have been interested in dictation as a test for learners of English as a foreign language. In the 1970s, dictation tests were seen as being useful tests of general language proficiency (Oller 1979). That is, they were not simply tests of listening skill, but drew on a more generalisable core of language knowledge. They were found to correlate highly with other tests and combinations of other tests, probably because they involved "active and creative processing" (Oller 1979: 41) which included a range of subskills involved in language use. Dictation involved normal contextual constraints on language, and required comprehension and to some degree production of meaningful sequences of language in relation to extralinguistic contexts (Oller 1979: 263). Dictation had the advantages of being a reliable test because of the large number of tested items it involved, of being an integrative test, and of correlating highly with other tests (Weir 1990). It was seen as having problems with marking, reliability of delivery, and possibly being a test of low level skills.

Typically a dictation test involves a passage around 150 words long, which is read aloud to the learners while they write what they hear. Usually the learners hear the whole passage once without breaks, then in phrases while they write what they hear, and then once again without breaks to check what they have written. Marking involves giving a point for each word written correctly, with some tolerance for minor spelling mistakes.

The test described in this article is a major adaptation of the traditional dictation test. The purpose of the adaptations was to develop a placement test which focused strongly on vocabulary knowledge, which could be used with learners of widely differing proficiencies, and which could be quickly and reliably marked by someone without special testing skills.

The reason for focusing on vocabulary knowledge was that as a placement test, it should focus on a very basic aspect of language knowledge that would affect a wide range of language skills. Vocabulary tests have been used to tap this kind of proficiency. Meara and Jones (1987 and 1990) developed a yes/no vocabulary test as a placement measure, and the Vocabulary Levels Test (Nation 1983 and 1990) has been used as an initial assessment tool for learners entering language programmes.

A placement test needs to be able to deal with learners of varying proficiencies in a sensible way. Such a test needs to have something that will stretch most learners, and something that most learners can do. Because a placement test needs to deal not only with learners who turn up on the first day of a course, but also learners who arrive late, it needs to be able to be administered and marked by anyone who is available to do so.

This article describes a type of dictation test which is easy to mark and which has four equivalent forms. These four forms of the test are included in the article. Each test takes about ten minutes.

A sample test

Let us look first at a script of the test. The instructions and the dictation are recorded on tape. Each test takes around twelve minutes.

Test A

This is a dictation test. You will hear a passage read once. Write what you hear. Remember, the passage will only be read once. It will not be repeated. Are you ready to begin writing?

Every year/a large number of young people/leave school and begin work./

Some obtain jobs on farms or in industry./ Others accept positions/in the government service./ Many seek posts in business or a trade./ A few with skills in art or music/ apply for work in these fields./

Their level of education frequently affects/ their range of possible openings./ Many firms, for instance,/ only select excellent candidates/ for training as future executives./ They will not consider applications from people/ with only average records of achievement at school./

What factors influence the choice of a career?/ The information available on this is uncertain/ but it is probable that finance,/ working conditions and prospects of improvement/ are the most significant considerations./ It seems apparent/ that organizations which retain their employees/ give them satisfaction in these respects./

A thorough investigation of the motives/ which operate in the selection of employment/ would prove a profitable topic for research./ Employers who would appreciate the assistance of the findings/ to enlist and maintain stable staff/ might be induced to invest in the project./

This is the end of the dictation test.

When learners sit the test they hear the text only once. At each pause, which is shown in the example by a slanting line (/), they write what they have just heard. When the test is marked however, the marker ignores the introductory paragraph and only gives points for the underlined words in the following four paragraphs. All other words and parts of the test are ignored.

Notice that the test has five sections. In each section some words have been underlined. In the first section of the test the underlined words are from the first 500 words of the Thorndike and Lorge (1944) word list. The underlined words in the second section are from the second 500 words of the Thorndike and Lorge list, in the third section from the second 1,000, in the fourth from the third 1,000, and in the fifth from the fourth to sixth 1,000. The range of difficulty in the test is thus primarily vocabulary based. The Thorndike and Lorge list is rather old, but it still is the only large lemmatized list. A check of the word levels against the Francis and Kucera (1982) study

which also uses lemmas showed that the frequency order of the words was similar between the two studies.

Notice also that the length of the dictated chunks, marked by a slanting line, increases from one section to another. In addition, each section is dictated slightly faster than the preceding section.

So, as the test moves from one section to another, the difficulty increases. In addition to an increase in vocabulary difficulty because of the decreasing frequency of the items, the length of chunks and the speed, there is also an increase in grammatical complexity. Table 1 summarizes some of the factors. This inclusion of a range of levels of difficulty means that parts of the test can be done by most learners and that there will also be parts that challenge them.

Table 1:
Design of the tests

	No. of key words in each paragraph	Frequency level of tested words	Average number of words (not just key words) per chunk
Introductory paragraph	10	1st 500	4.5
Paragraph 1	20	2nd 500	5.0
Paragraph 2	20	2nd 1,000	5.5
Paragraph 3	20	3rd 1,000	6.0
Paragraph 4	20	4th-6th 1,000	7.0

Using the test

The sample test given above and the accompanying instructions make it clear how the test should be presented. When marking, the following guidelines should be followed.

1 The introductory paragraph is not marked. Most people sitting the test get a large proportion of this section correct and it does not add to the discrimination of the test. It is used to give learners initial confidence.

2 Only the underlined words outside the introductory paragraph are marked. One mark is given for each of these words written correctly, so each of the four main paragraphs can earn 20 marks, making a possible maximum of 80 marks for the whole test.

3 Usually words are marked as correct if they are spelled correctly. Errors with the regular -s, -es, -d and -ed suffixes are ignored, and if the rest of the word is spelled correctly it is given a mark. The main principle to be followed is that there should be a consistent policy with regard to grammatical inflections and spelling. Ignoring spelling mistakes and insisting on correct inflections seems to yield similar results as long as the principle is followed consistently.

This marking procedure is easy to follow and after marking a few scripts can be very easily applied.

In the following sections of this article, we will look at the reliability and validity of the test, and determine if the four forms of the test included in this article are equivalent forms.

Reliability

The method of establishing reliability used in this study is that of equivalent forms. That is, the reliability of a test can be demonstrated by its high correlation with another test constructed in an equivalent way. The nature of the construction of the four forms has been described earlier in this article.

Table 2 shows that the tests correlated very highly with each other, with Pearson correlations in a very narrow range from .95 to .97. The tests are clearly reliable. Reliability is helped by the large number of points of assessment (the test is marked out of 80), by the tests being recorded, and by the focused marking system.

Table 2:
Inter-test correlations of the four forms

	B	C	D
A	0.97*	0.97*	0.97*
B		0.96*	0.96*
C			0.95*

* = all significant at the $p < .0000$ level

Validity

One way of seeing what the test measures is to look at it in terms of the various aspects of knowledge involved in knowing a word (Nation 1999). Table 3 presents a list of these aspects. The parts in italics are drawn on in the graded dictation test.

Table 3:
What is involved in knowing a word

Form	spoken	R P	*What does the word sound like?* How is the word pronounced?
	written	R P	What does the word look like? *How is the word written and spelled?*
	word parts	R P	What parts are recognizable in this word? What word parts are needed to express the meaning?
Meaning	form and meaning	R P	*What meaning does this word form signal?* What word form can be used to express this meaning?
	concept and referents	R P	*What is included in the concept?* What items can the concept refer to?
	associations	R P	What other words does this make us think of? What other words could we use instead of this one?
Use	grammatical functions	R P	*In what patterns does the word occur?* In what patterns must we use this word?
	collocations	R P	*What words or types of words occur with this one?* What words or types of words must we use with this one?
	constraints on use (register, frequency ...)	R P	Where, when, and how often would we expect to meet this word? Where, when, and how often can we use this word?

In column 3, R = receptive knowledge, P = productive knowledge.

It can be seen that the test measures mainly receptive knowledge, particularly recognition of the spoken form, being able to recall the meaning of the form, recognising the particular meaning it has in context, and recognising its typical collocates and the grammatical patterns it occurs in. The test also measures learners' productive knowledge of the written form, that is, its spelling. Because the test deals with vocabulary in use, it draws on a wide range of aspects of what is involved in knowing a word.

An item analysis of learners' scripts for the four tests provides evidence that some of these aspects of knowledge affected the difficulty of the tested words. Firstly, the fewer the number of words in a dictated chunk, the easier they were. The fewer the words, the easier it is to perceive and remember the forms. Secondly, the position of the tested word in a chunk affected its difficulty. In a chunk containing four test words, the last keyword was more likely to be written correctly than the other words. The first word was the next most likely, and the two middle words were the least likely. This suggests that grammatical and collocation cues may have been helping with the last word where there was opportunity for these cues to occur. In a chunk containing three test words, the first word was the one most likely to be written correctly. Thus, the number of keywords in a chunk is an important factor to control when making tests like these.

It was apparent that learners' failure to write a particular word could not be taken as evidence that learners did not know the word. The same word occurring in several tests often showed quite different difficulty indexes, usually as a result of its context. There were thirty-one examples of this, for example essential occurred in Test B and Test D. Factor occurred in all four tests. The factors affecting these different difficulty levels for the same word included position in the chunk (final position was easiest), number of target words in the chunk (the fewer, the easier), the length of the chunk containing the word (the shorter, the easier), and the grammatical function of the word. So, although knowledge of word form and meaning plays a significant role in the test, other contextual aspects can influence the result.

The test shows reasonable concurrent validity. Table 4 shows the results of a separate study looking at correlations between two of the dictation tests presented in this article and the Vocabulary Levels Test (Nation 1983 and 1990). The Vocabulary Levels Test measures vocabulary knowledge at five different word frequency levels which like the graded dictations were based mainly on the Thorndike and Lorge list. The reliability of the test is 0.94 (Read 1988).

Table 4:
Dictation and Vocabulary Levels Test correlations

	Dictation A	Dictation D
Vocabulary (n = 40)	0.78 p< 0.0001	0.78 p< 0.0001

The size of these correlations corresponds quite closely to correlations between other types of vocabulary tests. Different types of tests that focus on the same aspect of knowledge correlate with each other to a reasonable degree, but there is still a substantial amount of difference. Paul, Stallman and O'Rourke (1990), looking at native speakers of English, compared the three vocabulary test formats of multiple-choice, interview, and yes/no. They found reasonably high and significant correlations between the interview scores and the yes/no and multiple-choice scores, ranging from .66 to .81. The correlations show that the three kinds of tests are doing a similar job, but that there is enough unshared variance to see each of them as revealing some different aspects of vocabulary knowledge. Nist and Olejnik (1995) compared four different vocabulary tests of the same words, one requiring the native-speaking learners to write an illustrative sentence, another involving sentence completion, and others testing meanings and examples. The tests all correlated with coefficients of less than .7, showing that it is likely that different aspects of vocabulary knowledge were being tested, even though the same words were being tested. The dictation tests were largely testing different items, but these items were drawn from the same vocabulary frequency levels. The format and presentation of the tests was carefully controlled to be as similar as possible. These similarities account for the high correlations between the dictation tests. Their reasonably high correlation with the Vocabulary Levels Test can be accounted for by the fact that the Vocabulary Levels Test drew on similar vocabulary frequency levels from the same source, namely Thorndike and Lorge (1944).

Further work needs to be done on establishing the validity of the test as a placement test. Wall, Clapham, and Alderson (1994) describe the difficulties of evaluating a placement test. While it is relatively straightforward to gather evidence of internal reliability and validity, it is difficult to gain convincing external evidence against which a placement test can be compared. This external evidence includes gaining evidence of the effectiveness of the test in successfully placing learners in classes and in comparing scores on the placement test with appropriate external proficiency measures.

Equivalent forms

The appendix to this article contains three forms of the test in addition to the one presented earlier in this article. Henning (1987: 81-82) describes the standard procedures for developing equivalent forms, namely "to demonstrate equivalence of tests, these tests must (1) show equivalent difficulty as indicated by no significant difference in mean scores when the tests are administered to the same persons and their means are compared ..., (2) show equivalent variance ..., and (3) show equivalent covariance as indicated by no significant differences in correlation coefficients among equivalent forms or among correlation coefficients of equivalent forms with a concurrent criterion, all administered to the same persons". All four forms of the tests were administered to the same 45 learners over a four week period. The testing for equivalence also included the learners' score on a grammar test which was used as part of an entry test battery for a pre-university English course.

Table 5 shows that the means and standard deviations are very close to each other. Table 2 shows that the tests correlate extremely well with each other.

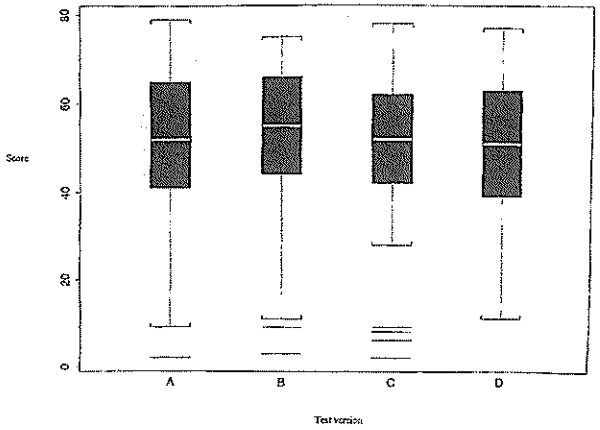
Table 5:
Means and standard deviations for the four forms

	Test A	Test B	Test C	Test D
Mean	50.29	51.47	49.73	49.02
s.d.	18.69	18.57	19.22	16.34

Figure 1 is a box plot showing the medians and quartiles of the tests. The Null Hypothesis of no significant difference between the means was

supported by a one-way ANOVA (F = 0.14, with 3 and 176 degrees of freedom, with p<.93), clearly indicating that the means were not significantly different.

Figure 1:
A box plot showing the medians and quartiles of the four versions



We have already seen (Table 2) that the four tests correlated very highly with each other. The covariances were tested using an approximate Chi-squared test. This returned a p value of .01. This suggests that the covariance matrix (Table 6) is not of the required sort with equivalent variances and covariances. However, the test is only approximate and more seriously it is extremely sensitive to outliers. The effect of these outliers can be seen particularly in the plot for C in Figure 1.

Table 6:
The covariances for the four test versions

	A	B	C	D
A	349.26	336.07	347.81	295.90
B		344.94	342.76	290.06
C			369.43	299.28
D				266.84

By removing the last four observations (all learners gaining scores of less than 20 out of 80 on each of the dictations), the covariances change. Table 7 presents the data re-analysed with the four very low scoring outliers removed, leaving a total of forty-one subjects

Table 7:
The covariances for the four test versions with outliers removed

	A	B	C	D
A	194.62	184.10	184.31	168.35
B		197.19	183.58	165.37
C			198.12	164.33
D				161.84

The corresponding ANOVA test for no difference in means gives an F value of .34 with 3 and 160 degrees of freedom. The corresponding p value is .8 so that under the Null Hypothesis of no difference an F value of .34 or greater would occur roughly 80% of the time, so there is no support for the alternative hypothesis of a difference in means.

This suggests that the tests will not be effective as equivalent forms for learners who get very low scores on the tests. For other learners the tests work very well indeed as equivalent forms and can be used as before-and-after tests to measure improvement during a course, or can be used interchangeably to maintain test security.

The dictation tests were correlated with a grammar test sat by the same subjects. All four versions correlated equally well with the grammar test (see Table 8).

Table 8:
Correlations of the four versions of the dictation test with a grammar test

	A	B	C	D
Grammar test	0.72*	0.70*	0.72*	0.73*

* significant at p<.01

The performance of the four tests indicates that they can be confidently regarded as being equivalent forms, particularly if learners performing very poorly on the tests are not included in the analysis.

Some final points

The dictation tests described in this article have features which make them attractive as a part of a placement test battery. They are reliable and valid tests of vocabulary knowledge, and they can discriminate at least as well as discrete point tests. The design of the tests allows them to be marked in a systematic way with fewer problems than are encountered in marking normal dictation. They can be administered in a short time. A disadvantage of the tests is because of the way they have been designed, the passages seem a little artificial. This fortunately does not seem to affect learner performance on the tests. Further research needs to be done on exploring the aspects of vocabulary knowledge that they test, and on validating their use as placement tests which is largely unexplored in this article.

Having four equivalent forms means that the tests can be used as measures of vocabulary growth using them as pre-tests and post-tests without being concerned about memory for the pre-test or any learning that it may have given rise to. As placement tests they can be used interchangeably to maintain test security.

Acknowledgments

We would like to express our gratitude to Dr Brian Dawkins of the School of Mathematical and Computing Sciences, Victoria University of Wellington for help with the statistical procedures and their interpretation.

References

Francis, W.N. and Kucera, H. 1982. *Frequency Analysis of English Usage.* Boston: Houghton Mifflin Company.

Henning, G. 1987. *A Guide to Language Testing.* Cambridge: Newbury House.

Meara, P. and G. Jones. 1987. Tests of vocabulary size in English as a foreign language. *Polyglot,* 8, Fiche 1.

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

Nation, I.S.P. 1983. Testing and teaching vocabulary. *Guidelines,* 5, 1, 12-25.

Nation, I. S. P. 1990. *Teaching and Learning Vocabulary.* Rowley, Mass.: Newbury House.

Nation, I.S.P. 1999. *Learning Vocabulary in Another Language.* ELI Occasional Publication No. 19, Victoria University of Wellington, New Zealand.

Nist, S.L. and S. Olejnik. 1995. The role of context and dictionary definitions on varying levels of word knowledge. *Reading Research Quarterly,* 30, 2, 172-193.

Oller, J.W. 1979. *Language Tests at School.* London: Longman.

Paul, P.V., A.C. Stallman, and J.P. O'Rourke. 1990. Using three test formats to assess good and poor readers' word knowledge. Technical Report No. 509 Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal,* 19, 2, 12-25.

Thorndike, E. L. and I. Lorge. 1944. *A Teacher's Word Book of 30,000 Words.* Columbia Teachers College.

Wall, D., C. Clapham, and J.C. Alderson. 1994. Evaluating a placement test. *Language Testing,* 11, 3, 321-344.

Weir, C.J. 1990. *Communicative Language Testing.* Hemel Hempstead: Prentice Hall.

APPENDIX

Test B

INTRODUCTORY PARAGRAPH

The demand for food/ becomes more important/ as the number of people in the world/ continues to increase./

PARAGRAPH 1

The duty to care/ for the members of a society/ lies with those who control it,/ but sometimes governments/ refuse to deal with this problem/ in a wise way,/ and fail to provide enough to eat./ When this occurs/ many ordinary people suffer./

PARAGRAPH 2

Often their economic situation/ does not permit them to create/ a system of regular supply./ When food is scarce,/ the pattern of distribution/ is generally not uniform./ In some areas production/ is sufficient to satisfy the needs of the population./ In others pockets of poverty exist./

PARAGRAPH 3

Using as their basis the research of experts/ to discover the factors/ in the previous failures to prevent starving,/ those in positions of leadership/ should institute reforms./ Unless ancient traditions of administration are overthrown/ the existence of the coming generations of mankind/ will be threatened./

PARAGRAPH 4

Though it is reasonable to presume that a reduction of consumption/ could be recommended in regions of prosperity,/ if this was enforced it would meet opposition/ with thousands rebelling/ in their determination to maintain their independence/ from those politicians dictating to them./ The selection of a differently devised procedure/ would be essential./

Test C

INTRODUCTORY PARAGRAPH

During the last twenty years/ many people left the country/ to work in the towns./

PARAGRAPH 1

Various reasons are advanced/ to explain this movement./ The use of machines on farms/ has reduced the demand for workers,/ while the growth of industry/ has increased the jobs in the cities./ This partly accounts for the change./

PARAGRAPH 2

In addition,/ the main centres offer/ a broad range of career possibilities,/ and the institutions providing higher education/ are generally located in the towns./ These factors tend to attract people/ who would otherwise be engaged in agriculture./

PARAGRAPH 3

But the total explanation of this drift/ is more involved./ A significant proportion of the population are employed/ in essential services to rural communities./ The shift to towns/ has decreased the necessity for these facilities./ In consequence, a reduction of people in these occupations/ has also occurred./

PARAGRAPH 4

This tendency, if the experts predict correctly,/ could result in a severe emergency./ Despite the obvious need to calculate approximately,/ the indications of the statistics of this transfer/ are that a serious crisis will shortly be inevitable/ arising from this irregular distribution of population.

Test D

INTRODUCTORY PARAGRAPH

My husband has a friend /who is the father/of a large family./ The poor man has ten sons./

PARAGRAPH 1

As a single man he imagined/ it would be a pleasant experience/ to bring up this number of children./ He now realizes/ it is a difficult problem/ to obtain enough money/ to provide food,/ clothes and shelter/ for a household of this size./

PARAGRAPH 2

Costs have tended to mount recently/ and parents frequently discover/ that unless they take special economic measures/ their expenses threaten to exceed their incomes./ If this occurs their debts compel them/ to make a reduction in their living standards./

PARAGRAPH 3

Besides financial considerations, another factor/ is that the effective running of a family of these proportions/ makes management ability essential./ Our acquaintance is fortunate to have a capable wife./ Before her marriage,/ she was employed in business administration./ She attributes her efficiency to this previous training./

PARAGRAPH 4

With his responsibility to maintain so many dependent relatives/ the man I refer to/ is continually striving to attain adequate security./ In his quest for promotion/ his slender resources are an unfortunate disadvantage,/ and furthermore his obligations hinder his prospects.