Available online at www.sciencedirect.com
ScienceDirect
System 35 (2007) 182–191
SYSTEM
www.elsevier.com/locate/system

Reliability assessment for two versions of Vocabulary Levels Tests

Peiling Xing [a,*], Glenn Fulcher [b]

[a] School of Humanities, CALS, University of Dundee, Dundee DD1 4HN, UK
[b] School of Education, University of Leicester, Leicester LE1 7RF, UK

Received 22 March 2006; received in revised form 25 October 2006; accepted 11 December 2006

Abstract

This article reports a reliability study of two versions of the Vocabulary Levels Test at the 5000 word level. This study was motivated by a finding from an ongoing longitudinal study of vocabulary acquisition that Version A and Version B of Vocabulary Levels Test at the 5000 word level were not parallel. In order to investigate this issue, Versions A and B were combined to create a single instrument. This was administered at one time to discover whether score differences found in the longitudinal study were present once the variable of time was removed. The data was analysed using correlation, and in order to discover if there was a significant difference between the two means of Version A and Version B, a *t*-test was used. Following that, a further examination of item facility values was conducted. The data analysis showed that Version A and Version B at the 5000 were highly correlated and highly reliable. However, the item analysis shows that the facility values of Version B contain a number of more difficult items. While versions of the Vocabulary Levels Tests at the 2000, 3000 and Academic levels may be treated as parallel for longitudinal studies, this does not hold at the 5000 word level. We suggest changes that need to be made to the test before it is used in future longitudinal vocabulary growth studies.

Keywords: Vocabulary Levels Test; Reliability

* Corresponding author.
E-mail addresses: p.xing@dundee.ac.uk (P. Xing), gf39@leicester.ac.uk (G. Fulcher).

0346-251X/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.system.2006.12.009

1. Introduction

The purpose of vocabulary assessment is to "monitor the learner's progress in vocabulary learning and to assess how adequate their vocabulary knowledge is to meet their communication needs" (Read, 2000, p. 2). There are two main branches of research in the field. One is testing vocabulary size and the other is measuring the quality of vocabulary knowledge. The former focuses on assessing the breadth of vocabulary while the latter assesses its depth. The present study is concerned with the assessment of vocabulary size and growth in longitudinal studies.

The participants' vocabulary sizes are assessed include native speakers and non-native speakers. Anderson and Freebody (1981, p. 96) present figures produced for native speakers in US universities, which range from 15,000 to 200,000 words. More research on L1 speakers can be seen in Thorndike (1924), Lorge and Chall (1963), and Nation (1993). Research on non-native speakers' vocabulary size mainly focuses on what minimum number of words international students need to know for the demands of their studies. Sutarsyah et al. (1994) estimate that knowledge of 4000–5000 English words will be a prerequisite for understanding an undergraduate economics textbook. Hazenberg and Hulstijin (1996) argue that a non-native speaker of Dutch in the first year at a university in the Netherlands needs a vocabulary of 10,000 Dutch words to be able to deal with reading materials. Read (2000, p. 83) argues that non-native speakers need to recognise at least 95% of the words in a text for efficient reading. Nation (1990) and Laufer (1992, 1997) argue that achieving at least the 3000 word level is necessary to meet this target. Read (2000) states that his two studies conducted in Indonesia show that first year learners fall short of this target. Because of the importance placed on vocabulary size as a measure of ability to cope with the demands of academic study, much research has been carried out to identify the vocabulary size of different groups of learners (Nation, 1983, 1990; Schmitt, 1993; Laufer, 1992, 1997; Laufer and Nation, 1995, 1999; Meara and Buxton, 1987; Meara and Jones, 1988, 1990; Meara, 1992, 1996). These studies use either Meara's Vocabulary Size Test, or Nation and Schmitt's Vocabulary Levels Test.

This paper investigates the Vocabulary Levels Test. It was compiled in the early 1980s by Paul Nation at Victoria University of Wellington in New Zealand. It was first used as a simple instrument for classroom use by teachers in order to help them develop a vocabulary teaching and learning programme. It was then published in Nation (1983, 1990) and has been widely used in New Zealand and many other countries. It is usually used to test the vocabulary size of migrant or international students when they first arrive in an English-speaking country. It has also been used by researchers who need an estimate of the vocabulary size of their subjects. Ten years later, Schmitt (1993) wrote three new forms of the test and took fresh samples of words for each level, following the original specifications. However, Schmitt used item facility values to remove items that did not discriminate between his students. The content of the current levels test is therefore not randomly selected, but is still claimed to be representative of the original levels. This new material was used by Beglar and Hunt (1999) who administered two forms each for the 2000-Word-Level and the University-Word-Level Tests to nearly 1000 learners of English in secondary and tertiary institutions in Japan. Based on the results, they selected 54 of the best-performing items to produce two new 27-item tests for each level. The two pairs of tests were then equated statistically. The authors concentrated on the two frequency levels. Schmitt has undertaken a similar test-development project with the four full forms of

the test. He administered the tests to106 non-native speaking British university students and created two longer versions which included 30 items instead of the original 18.

A range of well-known vocabulary test item types include *multiple-choice* for choosing the correct answer (Joe, 1994), *completion* for writing in the missing word (Read, 1995), *translation* for giving the L1 equivalent of the underlined word (Nurweni and Read, 1999), and *matching* format for matching each word with its meaning (see below). These item types are used in discrete, selective, context-independent vocabulary tests. The Vocabulary Levels Test used word–definition matching format to require test-takers to match the words to the definitions. Rather than giving a single estimate of total vocabulary size, it measures knowledge of words at five levels: 2000, 3000, 5000, 10,000, and academic English words. Each level contains 30 items. Table 1 displays a sample of three items as an independent unit. There are three definitions on the right and six words on the left. Candidates need to choose three out of the six words to match the three on the right.

In total at each level, 30 definitions need to be matched to 30 out of 60 words. Schmitt (personal communication, 2003) suggests the cutting point for the acquired level was 24. It means that if the candidate answered 24 (80%) questions correct, they had acquired the level. If not, it means they have not reached the level. While the basis for this assertion is not clear from published sources, it remains the basis for establishing vocabulary level in studies that use these tests.

In 2003/2004 academic year, the first author carried out a longitudinal study to investigate the growth in the vocabulary size of 52 Chinese learners at a UK university over one academic year. The instruments used in the study were Vocabulary Levels Test Version A (see Schmitt, 2000, pp. 192–200) and Version B (see Nation, 2001, pp. 416–424). They were treated as equivalent forms (Nation, 2001, p. 416). In order to decrease practice effect in the investigation, Version A and Version B were given alternatively at four points during the year, as shown in Fig. 1.

Based on the results of test 1, the 52 learners were divided into four sub-groups, as shown in Fig. 2. The description in details for the division rationale can be seen in Xing (2007).

Fig. 2 shows that the fifty two candidates were first divided into two groups (Group3000 and NGroup3000) depending on whether they reached the 3000 word level

Table 1
Three items for a unit of the Vocabulary Levels Test

a. Royal	
b. Slow	_____1____ The first
c. Original	_____2____ Not public
d. Sorry	_____3____ All added together
e. Total	
f. Private	

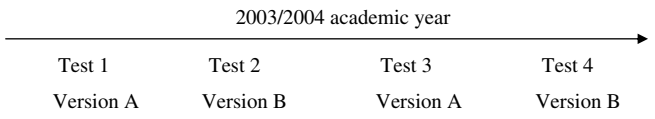

Fig. 1. The use of Version A and Version B to measure vocabulary growth.

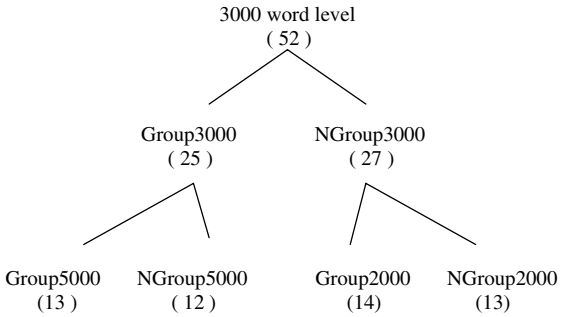

Fig. 2. The group division and number of candidates in each group.

or not (N stands for 'not'). Then based on whether they acquired 5000 or 2000, Group3000 was divided into Group5000 and NGroup5000, whereas NGroup3000 broke into Group2000 and NGroup2000. This creates four nominal groups which are NGroup2000 (learners who have not reached the 2000 word level), Group2000 (learners who have reached the 2000 word level, but have not reached the 3000 word level), Group3000 (learners who have reached the 3000 word level but have not reached the 5000 word level, or NGroup5000), and group5000 (learners who have reached the 5000 word level). Those four groups were abbreviated to GN2K, G2K, G3K and G5K, which are the terms used in the study.

The change of candidate numbers in different groups over one academic year is presented in Table 2.

Table 2 shows that with passing time the numbers in lower groups became smaller, while the numbers in higher groups increased. However, the pattern in NGroup5000 and Group5000 is very different. Group5000 is charted as: $13 \rightarrow 10 \rightarrow 15 \rightarrow 10$ and in NGroup5000 it is $12 \rightarrow 20 \rightarrow 16 \rightarrow 23$. It shows that the first and the third administration is one cluster and the second and the fourth administration is a second cluster. Since Version A was used in the first and the third test and Version B was used in the second and the fourth test, this raised the possibility that the results may be related to lack of reliability in one or both versions, or point to unequal difficulty levels.

The same problem was also found when the data at the 5000 word level was analysed using ANOVA. The data analysis method used was a two-way mixed-subject ANOVA. The results show there are significant main effects of group ($F(2, 34) = 62.714$; $p < .001$);

Table 2
The change of candidate numbers in different groups

Group	Number			
	Time 1	Time 2	Time 3	Time 4
NGroup2000	13	9	7	4
Group2000	14	12	14	15
NGroup3000	27	22	21	19
Group3000	25	30	31	33
NGroup5000	12	20	16	23
Group5000	13	10	15	10

but there is also a significant main effect of the test ($F(3, 102) = 18.074$; $p < .001$). This result is shown in Fig. 3.

Post hoc Bonferroni *t*-tests were used to look for the significance of comparisons among the tests. The results of the three groups NGroup2000, Group2000 and Group3000 showed the same trend. Version A, which was used for tests 1 and 3 appears to be easier than Version B, which was used for tests 2 and 4. Results for all other Vocabulary Levels Tests showed a steady rise in the scores across the four administration times, while the pattern for the 5000 word level test shows peaks and troughs. These are shown in Fig. 4 for three groups of learners.

Given the problem that had arisen during the longitudinal research, it became necessary to formulate a new research question to investigate the suitability of the instruments at this level.

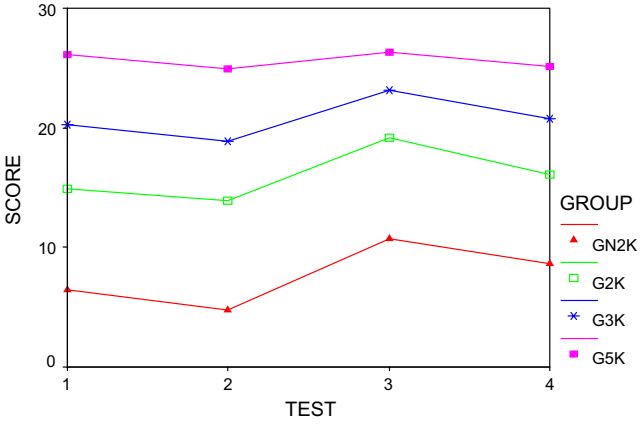

Fig. 3. Score trends at 5000 word level.

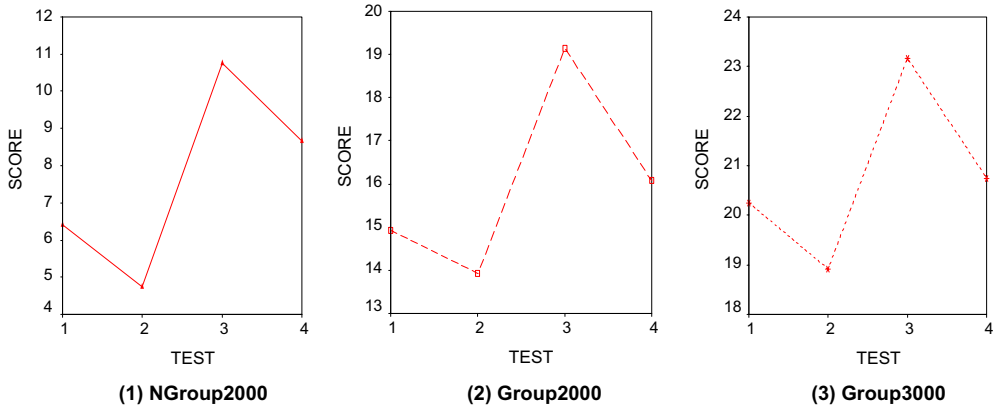

Fig. 4. Mean growth of NGroup2000, Group2000 and Group3000.

2. Research questions

Are Versions A and B of the 5000 word level test reliable, parallel forms?

3. Methods

3.1. Participants

Forty-six Chinese students who were newly-arrived in the UK took the two versions of the Vocabulary Levels Tests at one administration in 2005. Thirty percent of them were males and 70% were females. Their age ranges from 20 to 36 with the mean at age 23.31. The average duration of learning English was 10.55 years ranging from 7 to 24 years.

3.2. Instruments

The 5000 word level test Version A (Schmitt, 2000, pp. 196–197) and Version B (Nation, 2001, pp. 419–421) were combined to create an instrument that could be given in a single administration.

3.3. Procedure

In order to avoid fatigue or response order contamination, Version A and Version B was compiled into a single instrument by taking one item from Version A and then one item from Version B in sequence. Some items were used by both Version A and Version B. The common items only appear once in the mixed Version A + B.

Before the study was conducted, the mixed Version A + B was piloted on four Chinese students who were studying at the University of Dundee. The time needed to complete the test and difficulties they had in the process were observed, and the data used to set the time and administration condition for the main study.

3.4. Analysing the data

Pearson correlation was used to measure the strength or degree of an association between Version A and Version B. In order to find out whether the difference between the two versions is significant, a *t*-test was used. Examination of item facility values was then conducted in order to isolate sources of the difference between forms.

4. Results and discussion

The mean scores of Version A and Version B and their standard deviations are shown in Table 3.

The two versions were correlated using the Pearson product moment correlation, and the results of the analysis are summarised in Table 4.

The correlation coefficient is significant beyond the 1% level ($r = .844$; $n = 46$; $p < .001$), and we can conclude that the two versions of the test are linearly related, as can be seen in the scatter plot in Fig. 5.

Table 3
Mean scores and standard deviations (SDs) for Version A and Version B

Two Versions of Vocabulary Levels Test at the 5000 word level	Vocabulary Level Test	
	Mean	SD
Version A	17.46	6.93
Version B	15.07	7.88

Table 4
Summary table for Pearson correlation of Version A and Version B

Variables	Test	*r*	*p*	*N*
2 (Version A and Version B)	Two-tailed	.844[a]	.000	46

[a] Correlation is significant at the .01 level.

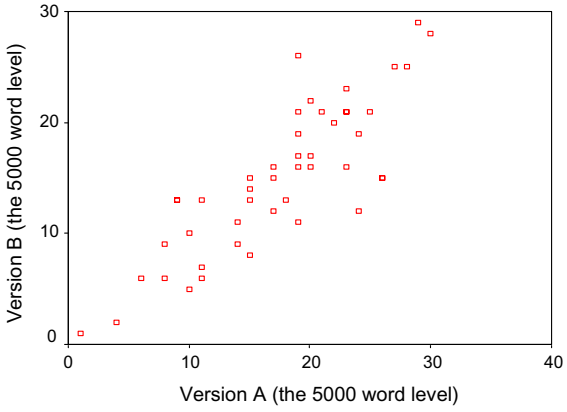

Fig. 5. Scatterplot of Version A against Version B.

Next, a *t*-test was conducted to investigate whether there was a significant difference between the mean scores on Version A and Version B. The results of the *t*-test ($t(45) = 4.251$; $p < .001$) confirmed that the two means are significantly different, as reported in Table 5.

Table 5 shows that Version B is somewhat more difficult than Version A, which can account for the pattern in Fig. 4.

Thus, a further investigation was carried out to examine the difficulties of the 60 items from Version A and Version B. The results are displayed in Table 6.

Table 6 shows that there were two difficult items in Version A which included two facility values smaller than .30 while there were six difficult items in Version B. These items are

Table 5
Descriptive scores of Version A and Version B

	Mean	Median	Mode	SD	Range	Mini.	Maxi.	*N*
Version A	17.5	19	19	6.9	29	1	30	46
Version B	15.1	15	13	6.7	28	1	29	46

Table 6
Item analysis of Version A and Version B

Version A			Version B		
Item	Number correct	Facility value	Item	Number correct	Facility value
A1	10	**.217**	B31	15	.326
A2	20	.435	B32	36	.783
A3	43	**.935***	B33	12	**.261**
A4	24	.522	B34	21	.457
A5	29	.630	B35	23	.500
A6	11	**.239**	B36	30	.652
A7	21	.457	B37	19	.413
A8	23	.500	B38	11	**.239**
A9	30	.652	B39	18	.391
A10	23	.500	B40	16	.348
A11	19	.413	B41	4	**.087**
A12	33	.717	B42	14	.304
A13	21	.457	B43	29	.630
A14	17	.370	B44	22	.478
A15	44	**.957***	B45	11	**.239**
A16	16	.348	B46	19	.413
A17	25	.543	B47	29	.630
A18	24	.522	B48	20	.435
A19	23	.500	B49	45	**.978***
A20	41	.891	B50	32	.696
A21	25	.543	B51	38	.826
A22	34	.739	B52	26	.565
A23	33	.717	B53	6	**.130**
A24	39	.848	B54	11	**.239**
A25	24	.522	B55	19	.413
A26	38	.826	B56	26	.565
A27	15	.326	B57	41	.891
A28	37	.804	B58	39	.848
A29	30	.652	B59	39	.848
A30	33	.717	B60	23	.500
Kuder Richardson 20 = .92			Kuder Richardson 20 = .90		

marked in the table by shading and bold numbers. A similar pattern is seen with regard to easier items. There are two facility values larger than .90 in Version A while there is only one in Version B (see the values in Table 6 marked by asterisk with bold numbers). However, Kuder Richardson 20 was over .90 for both versions, indicating that they are equally reliable.

It is therefore important to identify the sources of difficulty that make the two versions of the test non-parallel. The items identified in Table 6 are displayed in Table 7.

Table 7 shows that the more difficult words are rarely used in life. We subjected these items to analysis in Nation's Vocabulary Profiler, Web VP (1.0),[1] and each was recorded as being "off list" – or not frequent enough to appear in current word lists. However, the easy words in Table 7 are all very frequently used. Read (2000, p. 118) explains that the levels of the Vocabulary Levels Test "were defined by reference to the word-frequency

[1] http://www.er.uqam.ca/nobel/r21270/textools/web_vp.html.

Table 7
The difficult and easy items in Version A and Version B

Difficult items (facility value < .30)			
Word	Definition	Facility value	Item
Pail	Bucket	.217	A1
Apron	Cloth worn in front to protect your clothes	.239	A6
Gravel	Small stones mixed with sand	.261	B33
Stool	Seat without a back or arms	.239	B38
Creed	System of belief	.087	B41
Forge	Place where metals are made and shaped	.239	B45
Lurk	Hide and wait for someone	.130	B53
Resent	Feel angry about something	.239	B54
Easy items (facility value > .90)			
Balloon	Rubber bag that is filled with air	.935	A3
Document	A paper that provides information	.957	A15
Relax	Have a rest	.978	B49

data in Thorndike and Longe's (1944) list, with cross-checking against the General Service List (West, 1953) (for the 2000-word level) and Kucera and Francis (1967)". This list is clearly out of date with regard to the frequency of use of many lexical items today. At this level, the test designers may select some words that may have slipped out of the 5000 word list over the years.

In order to examine why the above items are difficult or easy for this population, the vocabulary teaching syllabus used in China was consulted, and it was found that all of the easy items (*balloon*, *document* and *relax*) belong to College English band-4 vocabulary which is a relatively lower level. As for the difficult items, *resent* is in Band-6 vocabulary level; *apron*, *gravel*, *stool*, and *creed* are in post-Band-6 vocabulary level; *pail* and *lurk* are not even in the syllabus. Only *forge* belongs to the Band-4 vocabulary level. But when it is checked in typical textbooks used for Band-4 level, it is not found in the books. According to the syllabus, Band-4 level is a basic requirement for college students while Band-6 level is an advanced requirement. This explains why these items are exceptionally difficult for Chinese students, and why older word frequency lists may be inappropriate tools for the selection of test items.

5. Conclusion

This study has examined and answered the question of why there was a problem with the 5000 word level vocabulary tests in a longitudinal study of vocabulary acquisition by Chinese learners in the UK University. The data analysis showed that Version A and Version B at the 5000 word level were highly correlated and highly reliable. But the item analysis has shown that Version B contained a number of harder items that mean the two versions of the test cannot be treated as parallel for research purposes. This study has identified a problem with the 5000 word level tests in their current format, and warns vocabulary researchers to take care in their use, especially in the context of longitudinal or gain scores studies. The tests are in need of some revision at the item level in order to relate score meaning to word frequency in current language use.

References

Anderson, R.C., Freebody, P., 1981. Vocabulary Knowledge. In: Guthrie, J.T. (Ed.), . In: Comprehension and teaching: research Reviews, vol. 2. International Reading Association, Newark: DE, pp. 77–117.

Beglar, D., Hunt, A., 1999. Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. Language Testing 16, 131–162.

Hazenberg, S., Hulstijin, J.H., 1996. Defining a minimal receptive second language vocabulary for non-native university students: an empirical investigation. Applied Linguistics 17, 145–163.

Joe, A., 1994. Generative use and vocabulary learning. Unpublished MA thesis, Victoria University of Wellington.

Kucera, H., Francis, W.M., 1967. A Computational Analysis of Present Day American English. Brown University Press, Providence, RI.

Laufer, B., 1992. How much lexis is necessary for reading comprehension? In: Arnaud, P.J.L., Bejoint, H. (Eds.), . In: Vocabulary and Applied Linguistics. Macmillan, London, pp. 126–132.

Laufer, B., 1997. The lexical plight in second language reading. In: Coady, J., Huckin, T. (Eds.), Second Language Vocabulary Acquisition. Cambridge University Press, Cambridge.

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

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

Lorge, I., Chall, J., 1963. Estimating the size of vocabularies of children and adults: an analysis of methodological issues. Journal of Experimental Education 32, 147–157.

Meara, P., 1992. EFL Vocabulary Tests. Centre for Applied Language Studies, University of Wales, Swansea.

Meara, P., 1996. The dimensions of lexical competence. In: Brown, G., Malmkjaer, K., Williams, J. (Eds.), Performance and Competence in Second Language Acquisition. Cambridge University Press, Cambridge, pp. 35–53.

Meara, P., Buxton, B., 1987. An alternative to multiple choice vocabulary tests. Language Testing 4, 142–154.

Meara, P., Jones, G., 1988. Vocabulary size as a placement indicator. In: Grunwell, P. (Ed.), Applied Linguistics in Society. Centre for Information on Language Teaching and Research, London, pp. 80–87.

Meara, P., Jones, G., 1990. Eurocentres Vocabulary Size Test, version E1.1/K10. Eurocentres Learning Service, Zurich.

Nation, I.S.P., 1990. Teaching and Learning Vocabulary. Heinle and Heinle, New York.

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

Nation, P., 1983. Testing and teaching vocabulary. Guidelines 5, 12–25.

Nation, P., 1993. Using dictionaries to estimate vocabulary size: essential but rarely followed procedures. Language Testing 10, 27–40.

Nurweni, A., Read, J., 1999. The English vocabulary knowledge of Indonesian university students. English for Specific Purposes 18, 161–175.

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

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

Schmitt, N. 1993. Forms B, C and D of the Vocabulary Levels Test. Unpublished manuscript.

Schmitt, N., 2000. Vocabulary in Language Teaching. Cambridge University Press, Cambridge.

Sutarsyah, C., Nation, P., Kennedy, G., 1994. How useful is EAP vocabulary for ESP? A corpus based case study. RELC Journal 25, 34–50.

Thorndike, E.L., 1924. The vocabularies of school pupils. In: Bell, J.C. (Ed.), Contributions to Education. World Book, Berlin, pp. 69–76.

Thorndike, E.L., Longe, I., 1944. The Teacher's Word Book of 30,000 Words. Teachers College, Columbia University, New York.

West, M., 1953. A General Service List of English Words. Longman, London.

Xing, P., 2007. Lexis Learning Rainbow: a longitudinal study of Chinese learners' vocabulary acquisition strategies in UK universities. Unpublished PhD thesis, University of Dundee, Dundee.